frontiers

Evaluating normalization accounts against the dense vowel space of Central Swedish

Anna S. Persson [1*], T. Florian Jaeger [2,3]

[1] Department of Swedish Language and Multilingualism, Stockholm University, Stockholm, Sweden
[2] Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA
[3] Computer Science, University of Rochester, Rochester, NY, USA

Correspondence*:
Anna S. Persson
anna.persson@su.se

ABSTRACT

Talkers vary in the phonetic realization of their vowels. One influential hypothesis holds that listeners overcome this inter-talker variability through pre-linguistic auditory mechanisms that normalize the acoustic or phonetic cues that form the input to speech recognition. Dozens of competing normalization accounts exist—including both accounts specific to vowel perception and general purpose accounts that can be applied to any type of cue. We add to the cross-linguistic literature on this matter by comparing normalization accounts against a new phonetically annotated vowel database of Swedish, a language with a particularly dense vowel inventory of 21 vowels differing in quality and quantity. All data and code for this study are shared via OSF, including the R markdown document that this article is generated from, and an R library that implements the models we present.

Keywords: vowel normalization; category separability; ideal observers; speech production; speech perception

INTRODUCTION

Talkers differ in their pronunciation of individual speech sounds due to both physiological differences and socio-cultural factors, including style, regional dialect, and second language accents. For listeners, this means that the mapping from acoustic cues to linguistic categories—phonemes, syllables, words, and ultimately word meanings—varies depending on the talker. How listeners manage to typically understand talkers despite this "lack of invariance" (Liberman et al., 1967) has remained one of the central questions for research on speech perception. Hypotheses about the mechanisms underlying this ability can be grouped into three, mutually compatible and complementary, accounts: (1) low-level, pre-linguistic auditory transformation of the acoustic signal, (2) learning of changes in the linguistic representations, and (3) post-linguistic changes in decision-making biases (see e.g., Johnson, 2006; Pardo and Remez, 2006; Xie et al., 2022). The present study focuses on the first type of account, that the acoustic signal is transformed and normalized early on during auditory processing (for recent reviews, Johnson and Sjerps, 2021; Stilp, 2020).

The existence of some form of pre-linguistic normalization is motivated by *a priori* considerations about the physics of sounds (cf. the discussion of uniform scaling in Barreda, 2020), evolutionary arguments (e.g., even non-human animals exhibit similar abilities, as reviewed in Barreda, 2020), as well as brain imaging evidence: talker-normalized information about the speech signal can be decoded from areas as early as the brain stem (e.g., Sjerps et al., 2019; Sussman, 1986), and thus prior to even the earliest cortical areas typically associated with linguistic category representations or decision-making. While it is rather uncontroversial that normalization is part of adaptive speech perception, questions remain about the specific nature of the operations involved in normalization. We contribute to this line of research by comparing different types of normalization accounts against data from the production of short and long vowels of Central Swedish.

Normalization accounts were originally proposed as a theory of how the brain removes *physiologically*-conditioned variation from the speech signal, reducing variability in, for example, category means between talkers, and thus reducing the overlap of phonological categories in the acoustic-phonetic space (e.g., Bladon et al., 1984; Gerstman, 1968; Lobanov, 1971; Miller, 1989; Nearey, 1978; Nordström and Lindblom, 1975; Peterson, 1961; Syrdal and Gopal, 1986; Sussman, 1986). Most of this early work focused specifically on differences in formants, which cross-linguistically are the primary cues to distinctions in vowel quality, and known to be affected by the vocal tract size of the talker (e.g., Fox et al., 1995; Peterson and Barney, 1952; Verbrugge and Shankweiler, 1977; Yang and Fox, 2014). Figure 1 illustrates the effect of one of the most commonly applied normalization accounts (Lobanov, 1971) for the vowel spaces of three L1 speakers of US English (from the database reported in Xie and Jaeger, 2020). By normalizing cues prior to categorization, physiological differences between talkers can be reduced, resulting in reduced between-talker variability (compare Figure 1B to 1A). If listeners' category representations pool experiences across talkers into a single talker-independent model, this reduced inter-talker variability results in reduced category overlap (compare Figure 1D to 1C).[1]

[1] Talker-independent category representations are assumed in many influential models of spoken word recognition (e.g., Luce and Pisoni, 1998; McClelland and Elman, 1986; Norris and McQueen, 2008). While talker-independent representations might be a simplifying assumption for some of these theories, this assumption has persisted for decades (e.g. Magnuson et al., 2020; ten Bosch et al., 2022). Exceptions include exemplar accounts (e.g., Johnson, 1997; Pierrehumbert, 2001) and the Bayesian ideal adaptor account (Kleinschmidt and Jaeger, 2015). Importantly, it is an unresolved question whether—or for which cues and phonetic contrasts—listeners maintain talker-specific *category* representations (for findings and discussion, see Kraljic and Samuel, 2007; Kleinschmidt and Jaeger, 2015; Kleinschmidt, 2019; Xie et al., 2021). In the present paper, we follow previous work and compare the effectiveness of normalization under the assumption of talker-independent category representations.

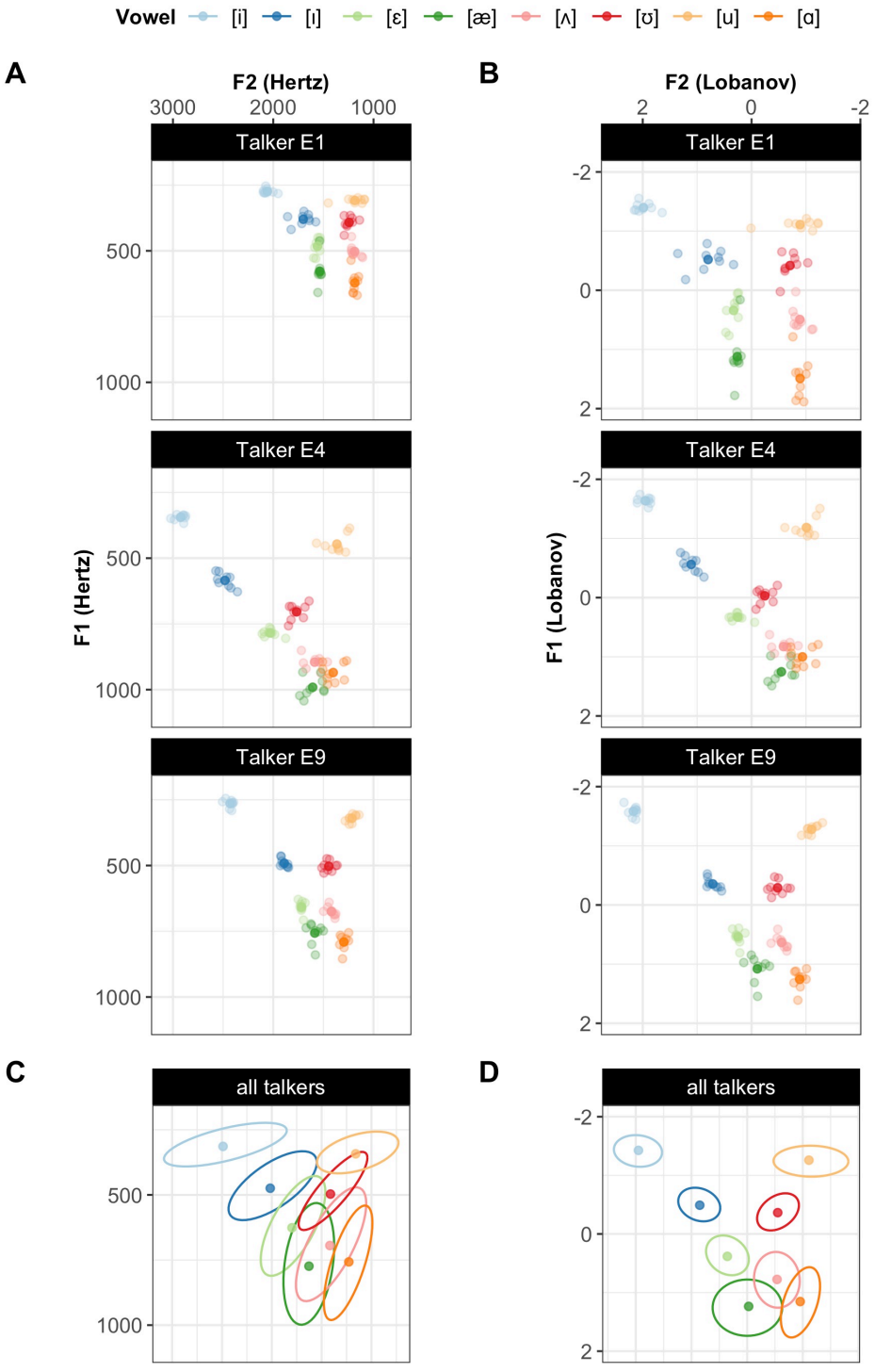

Figure 1. Illustrating how normalization reduces category overlap for the 8 monophthongs of L1 US English. Three talkers from the Xie and Jaeger (2020) database are shown before (**Panel A**) and after Lobanov normalization (**Panel B**). Lobanov normalization reduces inter-talker variability in the category means and, to some extent, in the category variances. The bottom two panels aggregate the data from all 17 talkers in the database (5 female, 12 male), showing the means and 95% probability mass bivariate Gaussian densities for each vowel before (**Panels C**) and after Lobanov normalization (**Panel D**).

Dozens of different accounts of vowel normalization have been proposed over the years (e.g., Bladon et al., 1984; Fant, 1975; Gerstman, 1968; Joos, 1948; Lobanov, 1971; Miller, 1989; Nearey, 1978; Nordström and Lindblom, 1975; Syrdal and Gopal, 1986; Traunmüller, 1981; Watt and Fabricius, 2002; Zahorian and Jagharghi, 1991; for reviews, see Barreda, 2020; Weatherholtz and Jaeger, 2016). Carpenter and Govindarajan (1993) summarize over 100 different vowel-specific normalization accounts, many of them closely related to each other. More recently, additional *general* normalization accounts have emerged that can be applied to *any* type of cue and phonological contrast (e.g., Cole et al., 2010; McMurray and Jongman, 2011). The most widely used of these proposals, C-CuRE, has since been successfully applied to the categorization of US English fricatives (Apfelbaum et al., 2014; Crinnion et al., 2020; McMurray and Jongman, 2011), stop voicing (Kulikov, 2022; Toscano and McMurray, 2015; Xie et al., 2022), vowels (Kleinschmidt, 2019; McMurray and Jongman, 2016), and sentence-final rising question vs. statement intonation (Xie et al., 2021). In each of these studies, C-CuRE reduced inter-talker variability and improved categorization. C-CuRE stands for "**c**omputing **cu**es **r**elative to **e**xpectations", capturing the motivation behind many of the earlier normalization accounts that cue values should be interpreted relative to the distribution they are expected to have in the present context. In addition to being more general than earlier accounts, C-CuRE also addresses one of the main arguments against normalization as an account of human speech perception (e.g., Johnson, 2006, which continues to be frequently cited in reviews of exemplar theory): by focusing on listeners' expectations rather than talkers' physiology, accounts like C-CuRE capture that inter-talker variability is not limited to physiology.

Table 1 lists the normalization accounts investigated in the present study. This includes both the most influential vowel-specific normalization accounts that have been found to perform well in previous works (e.g., Lobanov and Nearey2 normalization) and several variants of the general purpose normalization C-CuRE. As indicated through shading in the table, the accounts can be grouped into four types based on the computational assumptions they make. *Transformations* are meant to transform the formant data from acoustic (Hz) into a perceptual space that approximates the perceptual organization of auditory information in the human brain. All other accounts instead or additionally adjust each formant value based on either the values of other formants on the same segment (*vowel-intrinsic* approaches) or summary statistics of the formant across segments (*vowel-extrinsic* approaches).[2] We further distinguish two types of vowel-extrinsic approaches that differ in their computational complexity and tractability: approaches that *center* each cue relative to its mean across all vowel segments, and approaches that instead/additionally *standardize* cues relative to the overall variability or range of the cue across all vowel segments (for reviews, see also e.g., Johnson, 2005; Kohn and Farrington, 2012; Weatherholtz and Jaeger, 2016).[3] The former type includes C-CuRE, and we consider different variants of this approach, one for each transformation approach in Table 1.

Our selection of accounts to consider in the present study is primarily based on their influence and performance in previous evaluations against other data sets. Additionally, we only consider

[2] Miller's formant-ratio account (Miller, 1989) is technically a hybrid approach: the first formant (F1) is normalized with regard to an extrinsic sensory reference (based on the average F0 across segments); subsequent formants are (intrinsicly) normalized using the normalized lower formants on the same vowel segment.

[3] Here we group accounts based on their computational complexity (the number of parameters listeners are assumed to estimate). For example, we group Nearey1 and Nearey2 with the centering accounts because they require estimation of only cue means. However, since these accounts perform centering over log-transformed Hz, they can also be considered as a form of functionally constrained standardization in non-log space (Barreda and Nearey, 2018).

accounts that are sufficiently general in nature to be applied across languages. This decision stems from our goal to understand the mechanisms underlying *human* speech perception. This means that we for instance do not include Watt & Fabricius (Watt and Fabricius, 2002; Fabricius et al., 2009), another frequently used normalization account, as it requires making specific assumptions of vowel inventories of the language. Finally, we do not consider *combinations* of accounts. This follows the majority of previous work but is an important limitation that we return to in the general discussion.

Table 1. Normalization accounts considered in the present study. Unless otherwise marked, formant variables (Fs) in the right-handside of normalization formulas are in Hz.

		Normalization procedure	Perceptual scale	Source	Formula
		n/a	Hz	n/a	n/a
trans-formation		n/a	Bark	Traunmüller (1990)	$F_n^{Bark} = \frac{26.81 \times F_n}{1960 + F_n} - 0.53$
		—	ERB	Glasberg & Moore (1990)	$F_n^{ERB} = 21.4 \times \log_{10}(1 + F_n \times 0.00437)$
		—	Mel	Stevens & Volkmann (1940)	$F_n^{Mel} = 2595 \times \log_{10}(1 + \frac{F_n}{700})$
		—	Semitones conversion	Fant et al. (2002)	$F_n^{ST} = 12 \times \frac{ln(\frac{F_n}{100})}{ln}$
intrinsic		Syrdal & Gopal's Bark-distance model[4]	Bark	Syrdal & Gopal (1986)	$F1^{SyrdalGopal} = F1^{Bark} - F0^{Bark}$ $F2^{SyrdalGopal} = F2^{Bark} - F1^{Bark}$
		Miller (formant-ratio)	log	Miller (1989)	$SR = k(\frac{GMf0}{k})^{1/3}$ $F1^{Miller} = log(\frac{F1}{SR})$ $F2^{Miller} = log(\frac{F2}{F1})$ $F3^{Miller} = log(\frac{F3}{F2})$
extrinsic	centering	C-CuRE	Hz	McMurray & Jongman (2011)	$F_n^{C-CuRE} = F_n - mean(F_n)$
		—	Bark		
		—	ERB		
		—	Mel		
		—	Semitones conversion		
		Nearey1 (log-mean)	log	Nearey (1978)	$F_n^{Nearey1} = \ln(F_n) - mean(ln(F_n))$
		Nearey2 (single parameter log-mean)	log	Nearey (1978)	$F_n^{Nearey2} = \ln(F_n) - mean(ln(F))$
	standardizing	Gerstman (range normalization)	Hz	Gerstman (1968)	$F_n^{Gerstman} = 999 \times \frac{F_n - F_n^{min}}{F_n^{max} - F_n^{min}}$
		Lobanov (z-score)	Hz	Lobanov (1971)	$F_n^{Lobanov} = \frac{F_n - mean(F_n)}{sd(F_n)}$

[4] Previous work has considered two different implementations of Syrdal & Gopal's Bark-distance model for F2, depending on the language (Adank, 2003; Fant, 1983; Syrdal & Gopal, 1986). In the SI (see Section Evaluation of implementations of Syrdal & Gopal), we compare these two implementations, and find that the F2-F1 implementation performs better for the present data. We thus present that version of Syrdal & Gopal's model in the main text of Studies 1 and 2.

Previous comparisons of normalization accounts have primarily focused on English (e.g., Adank et al., 2004; Barreda and Nearey, 2018; Carpenter and Govindarajan, 1993; Clopper, 2009; Disner, 1980; Escudero and Bion, 2007; Fabricius et al., 2009; Flynn and Foulkes, 2011; Hindle, 1978; Kohn and Farrington, 2012; Labov, 2010; Richter et al., 2017; Syrdal, 1985). Additional studies have investigated, for example, Dutch (Adank et al., 2004; Disner, 1980), Russian (Lobanov, 1971), and Brazilian Portuguese (Escudero and Bion, 2007). The complexity of the vowel inventories (7-11 monophthongs) and the number of these vowels included in the comparison (2-11) varied across these studies. We add to this literature by comparing normalization accounts against a new phonetically annotated database of Central Swedish (SwehVd, introduced below). With a total of 21 monophthong allophones that vary in quantity (long vs. short vowels) and quality, the vowel inventory of Swedish is crowded compared to most languages previously studied in the normalization literature. This allows us to test whether the same normalization accounts that work well for simpler vowel inventories generalize well to more crowded vowel spaces. Additionally, the presence of quantity contrasts between long and short allophones means that Swedish provides a suitable case study to bridge the literature between vowel-specific normalization accounts (which focus on formants, and thus only quality contrasts) and general normalization accounts that can be applied to any type of cue (and thus also vowel duration, which is expected to be the primary cue to vowel quantity). Relatedly, previous studies have found both F3 and vowel duration to be important cues to vowel categorization in Swedish (e.g., Behne et al., 1997; Fujimura, 1967; Hadding-Koch and Abramson, 1964). These two cues have never (duration) or rarely (F3, but see e.g. Adank et al., 2004; Barreda and Nearey, 2018; Carpenter and Govindarajan, 1993; Nearey, 1989; Syrdal, 1985) been included in comparisons of normalization accounts.

We compare the normalization accounts in Table 1 both in terms of their effectiveness in reducing within-category variability relative to between-category variability (Study 1) and in terms of the predicted consequences for recognition accuracy (Study 2). We originally only conducted Study 2, as it more adequately addresses our goal of assessing the predicted consequences of normalization for perception. However, we later added Study 1 because—despite serious shortcomings—measures of between- vs. within-category separability/variability continue to be commonly used in research on normalization. The primary purpose of including Study 1 is to illustrate some of these shortcomings, highlighting the benefits of using models of speech perception when evaluating normalization accounts (as we do in Study 2, for related discussion, see Barreda, 2021).

Following the bulk of previous work (but see e.g., Barreda, 2021; McMurray and Jongman, 2016; Nearey, 1989; Richter et al., 2017), the present studies do not compare the predicted consequences of normalization against categorization responses from human listeners, though we plan to do so in future work. We discuss this and other limitations—most shared with previous work—after presenting our studies. As part of that discussion, we also raise conceptual considerations about the *a priori* plausibility of different normalization accounts.

Both Study 1 and 2 compare normalization accounts applied to (1) only F1 and F2, as in the majority of previous studies, (2) F1-F3, as in, e.g., Adank et al. (2004), and (3) F0-F3 as well as vowel duration. This allows us to assess whether differences in the effectiveness of normalization accounts depend on the number and types of cues that are considered. Since listeners integrate cues beyond F1 and F2 (e.g., Assmann et al., 1982; Hillenbrand and Nearey, 1999; Nearey and Assmann, 1986), this is an important gap in evaluating the plausibility of different normalization accounts as models of adaptive speech perception. All three comparisons are evaluated both separately for

short and long vowels, and for the entire space of the 21 vowels. This allows us to assess whether the same types of normalization perform well across the entire vowel inventory.

To the best of our knowledge, only one previous study has compared normalization accounts against Swedish, as part of a cross-linguistic comparison across six Germanic languages (Disner, 1980). Disner (1980) compared 4 normalization accounts, using F1 and F2 means of the nine long Swedish vowels spoken by 24 male Swedish talkers (from a database presented in Fant et al., 1969). Of interest to the present study, the results for Swedish differed from the other Germanic languages in two unexpected ways. Whereas Lobanov normalization—which involves centering and standardizing (cf. Table 1)—performed best for Swedish, Nearey2 normalization—which involves only centering—performed best for the other four languages. And, while normalization effectively reduced inter-talker variability in category variances for the other four languages by 61%-71%, it was substantially less effective for Swedish (41%). As discussed by Disner (1980), this raises the question as to whether these findings reflect an inherent property of Swedish or merely differences in the phonetically annotated databases available for each language. In particular, the Swedish data consisted of *vowels* produced in isolation without any lexical or phonetic context, whereas the data for the five other languages consisted of isolated *word* productions (paralleling the majority of research on normalization). The present study addresses this difference: the new database we introduce consists of *h*-VOWEL-*d* word recordings, which makes our stimuli directly comparable to those used in previous work on normalization. Additionally, we complement Disner's study by focusing on female, rather than male talkers, by considering both long and short vowels (separately and together), and by including the general normalization account C-CuRE. This lets us revisit whether simple *centering* accounts perform best for Swedish—like for the other languages in Disner (1980).

All data and code for this article can be downloaded from OSF at `https://osf.io/zb8gx/`. This article is written in R markdown, allowing readers to replicate our analyses using freely available software (R Core Team, 2021; RStudio Team, 2020), while changing any of the parameters of our models. Readers can revisit the assumptions we make—for example, by substituting alternative normalization models. The supplementary information (SI) lists the software/libraries required to compile this document. Next we introduce the new phonetically annotated corpus of Central Swedish vowel productions employed in the present studies.

THE SWEHVD DATABASE

The SwehVd database is a new phonetically annotated corpus of Swedish *h*-VOWEL-*d* (short: hVd) word recordings. All recordings, annotations, and acoustic measurements are available on OSF at `https://osf.io/ruxnb/`. SwehVd was collected with the goal to characterize the Central Swedish vowel space—specifically, the Stockholm variety—within and across talkers. It covers the entire monophthong inventory of Central Swedish, including all nine long vowels (*hid*, *hyd*, *hud*, *hed*, *häd*, *höd*, *had*, *håd*, *hod*), eight short vowels (*hidd*, *hydd*, *hudd*, *hedd*, *hädd*, *hödd*, *hadd*, *hådd*, *hodd*), and four allophones (*härd*, *härr*, *hörd*, *hörr*). To our knowledge, there are few publicly available databases of Swedish vowel productions that are phonetically annotated (e.g., Bruce et al., 1999; Eklund and Traunmüller, 1997; Fant et al., 1969; Kuronen, 2000). The largest and perhaps best-known is SweDia 2000 (Bruce et al., 1999). SweDia 2000 was developed to characterize differences in vowel pronunciations *across* regional varieties of Swedish. It consists of recordings of spontaneous speech, isolated words in varying phonological contexts, and phrases in isolation from approximately

1300 talkers of 107 regional backgrounds, with 10-12 recorded talkers per region and 5-15 recordings per vowel for each talker.

Unlike most existing databases, SwehVd focuses on a single regional variety, providing high resolution within and across talkers for this variety: SwehVd consists of N=10 recordings of each hVd word (for a total of 210 recordings for the 21 different hVd words) per talker. We note that this makes the demographic composition of SwehVd relatively homogenous, compared both to some other vowel corpora of Swedish (e.g., Bruce et al., 1999; Kuronen, 2000) and to some previous studies on normalization—a point we return to in the general discussion. Specifically, we target N = 24 male and female talkers each (current N = 23, all female) for a total targeted N of tokens = 10,080 (current N = 4511 tokens). The accompanying database contains first to third formant (F1-F3) measurements for each talker at five time points across each vowel, together with vowel duration and mean F0 over the entire vowel.

SwehVd follows the gross of research on normalization and uses hVd words for recording in order to minimise coarticulatory effects from the surrounding phonetic context. The hVd context was originally chosen for studies on English because the glottal /h/ in onset position minimizes supraglottal articulations (confirmed in e.g., Chesworth et al., 2003; Robb and Chen, 2009). Since then hVd words have played a central role in research on vowel production (e.g., Hillenbrand et al., 1995; Peterson and Barney, 1952) and perception (e.g., Malinasky et al., 2020; Peterson and Barney, 1952). Since Swedish onset /h/ is a glottal approximant (Riad, 2014) similar to English, the use of this context in SwehVd facilitates comparison to similar databases from other languages. It deviates, however, from the majority of previous studies on Swedish vowels, which have either not held phonetic context constant across vowels (e.g., Bruce et al., 1999), or have investigated vowel production out of context (Eklund and Traunmüller, 1997; Fant et al., 1969; Disner, 1980) or in different CVC contexts (e.g., *k*V*p* and *p*V*k* in Nordstrand et al., 2004; *v*V*t*, *v*V*tt*, *f*V*t*, *f*V*tt*, in Behne et al., 1997).

The Swedish vowel inventory

The Central Swedish vowel inventory contains 21 monophthong vowels. Seventeen of these vowels form nine pairs distinguished by quantity (long and short): in Central Swedish, the two long vowels [ɛː] and [eː] both neutralize to the same short vowel [ɛ] (resulting in a total of 17, rather than 18, distinct vowels). The two variants of a pair are considered allophones, the selection of which is determined primarily by stress and syllable complexity. Quantity is neutralized in unstressed positions (Riad, 2014).[5] Vowels lengthen in open word-final syllables, before morpheme-final single consonants, and in non-final syllables.

Additionally, there are four contextually conditioned allophones to [ɛ] and [ø]. Before /r/ (or any retroflex segment), both the long and short versions of these vowels lower to long and short [æ] and [œ], respectively. As shown in Table 2 (adapted from Riad, 2014), some long-short vowel pairs are described to differ not only in quantity but also in quality: generally, short vowels are described as more open and also more centralized, forming a more condensed vowel space. In ongoing work (Persson, 2023), we found this to be confirmed for SwehVd.

Several of the long vowels have been claimed to be diphthongized in Central Swedish (e.g., Elert, 1981; Fant et al., 1969; Fant, 1971; Kuronen, 2000) and/or with consonantal elements (McAllister

[5] This reflects the mainstream analytical position in present-day Swedish phonology. The opposite position, distinctive vowel quantity, has also been proposed (e.g., Linell, 1978, 1979; Schaeffler, 2005). This theoretical debate does not affect the interpretation of our results.

Table 2. The phonological characterization of long (left) and short (right) Central Swedish vowels (based on Riad, 2014)

	front	rounded	central	back
high	[iː]	[yː]		[uː]
mid-high	[eː]	[ʉː]		[oː]
mid	[ɛː]	[øː]		
low	[æː]	[œː]		[ɑː]

	front	rounded	central	back
high	[ɪ]	[ʏ]	[ɵ]	[ʊ]
mid	[ɛ]	[ø]	[œ]	[ɔ]
low	[æ]		[a]	

et al., 1974), though empirical evaluations of this claim have returned mixed results (Eklund and Traunmüller, 1997; Fant et al., 1969; Leinonen, 2010). In separate work (Persson, 2023), we have found evidence for diphthongization of some vowels in SwehVd, including some for which it has not previously been reported. Here we do not discuss this issue further since it has no direct consequences for the present study: normalization can be applied to both monophthongs and diphthongs, and listeners presumably use the full information contained in the dynamics of formant trajectories—rather than just aggregate point estimates (the simplifying information we make below, following other work on normalization).

Participants

Stockholm Swedish is a variety subsumed under Central Swedish—the regional standard variety of Swedish spoken in an area around and beyond Stockholm (eastern Svealand), including Mälardalssvenska, Sveamål, Uppsvenska, Mellansvenska (see e.g., Bruce, 2009; Elert, 1994; Riad, 2014). SwehVd targets the contemporary Central Swedish spoken in the larger Stockholm area.

Native talkers of Stockholm Swedish were recruited through word-of-mouth, flyers at Stockholm University Campus (see example flyer in SI, Participant recruitment), and online channels (accindi.se). Participants were selected based on the following criteria: L1 talkers of Swedish, born and raised in the greater Stockholm area or its surroundings, 20-40 years old (mean age=28; SD=5.35). Four of the participants were bilingual from birth, the L1s of each talker are provided in the database. All participants were reimbursed with a voucher to the value of SEK 100 after completing the recordings.

Recording procedure

Recording for the SwehVd database began in 2020 and is ongoing. The data were collected by the first author and Maryann Tan (Stockholm University). The hVd words were recorded together with another set of recordings targeting the production of Swedish word-initial stop voicing. Recording took place in a sound-attenuated room at the Multilingualism Laboratory, Department of Swedish Language and Multilingualism, Stockholm University.

Prior to recording, participants were informed about the study and given the possibility to ask questions before signing a consent form. They were then given instructions and seated at approximately 10 cm distance from an Audio Technica AT3035 microphone facing a computer screen. Words were presented one at a time, centered on screen, using PsychoPy software (Peirce et al., 2019). Participants were instructed to read the words with their natural voice as they appeared on screen. Each talker read the same 21 target words, with 48 mono- and bi-syllabic filler words interspersed. Each target word was repeated 10 times and each filler word was repeated five times, generating a total of 450 productions per talker, 210 target productions and 240 filler productions. We generated two pseudo-randomized lists of the words, each list divided into four different blocks. Words were blocked across block lists and randomized within block lists, with the constraint that

the same word would not appear more than twice in succession. Each participant was randomly assigned to one of the two lists. The pace of the presentation of the words was controlled by the experimenter, who was listening over Sennheiser HD215 headphones in the next room. A Yamaha MG102c mixing console with a built-in preamplifier was used together with a high-end ground isolator for preventing signal interference (Monacor FGA-40HQ). The speech was recorded at 44.1 kHz in Audacity (Audacity, 2021). Each long sound file was split into individual short sound files of one word each. The boundaries of each file were slightly trimmed and the files were labelled with the target word. All sound files from the same talker were concatenated into one long file before further processing.

The complete list of target hVd words is provided in Table 5 in the SI. It consists of four real Swedish words, *hed*, *härd*, *hörd*, *hud* (English translations: *heath*, *hearth*, *heard*, and *skin*, respectively) and 17 phonotactically legal pseudowords. Following Swedish orthographical conventions for quantity, we used orthographic *hVdd* to elicit the short vowel allophone (e.g., *hudd* for [ɵ]) and orthographic *hVd* to elicit the long vowel allophone (e.g., *hud* for [ʉː]). This orthography reflects systematic phonological process of complementary quantity in Swedish (Riad, 2014). In order to elicit the contextual allophones to [ɛ] and [ø], we added the supradental [ɖ] to elicit the long allophones (*härd*, *hörd*), and [ɾ] to elicit the short allophones (*härr*, *hörr*). In a small-scale pilot preceding recordings, the expected transparency of the orthography for eliciting the long and short vowels was confirmed by three native talkers and one non-native talker of Swedish (these talkers did not participate in the study). However, *hodd* [ʊ] and *hod* [uː] sometimes elicited [ɔ].[6] We therefore decided to add instructions to the participants for these two words. When *hod* or *hodd* appeared on screen, a written guide indicating the target vowel appeared below the word in smaller font size: "hod som i hot", "hodd som i hosta", with *hot* and *hosta* being real Swedish words containing [uː] and [ʊ], respectively.[7] Whenever the experimenter noticed that the pronunciations clearly targeted another vowel, recordings were stopped and participants were reminded to carefully read the guide.

The recordings were divided into five blocks: one practice block and four recording blocks, with breaks in between. The purpose of the practice block was threefold: to familiarize the participants with the recording procedure, to adjust the recording level, and if necessary, to further instruct the participant (e.g., if the participant used inappropriate or inconsistent intonation or stress pattern). Each recording block consisted of either 110 (N=2 blocks) or 120 (N=2 blocks) trials. The length of each block was approximately eight minutes, for a total of roughly 30 minutes recording time per talker. After the recording, participants filled out a language background questionnaire and received their reimbursement.

Word and vowel segmentation

SweFA, a Swedish version of the Montreal Forced Aligner developed by Young and McGarrah (2021), was used to obtain estimates for word and segment boundaries. The boundaries were manually corrected by the first author (an L1 talker of Central Swedish). Following standard segmentation protocol and guidelines in Engstrand et al. (2001), segment boundaries were adjusted using spectrogram, waveforms and pitch and intensity tracks. The boundaries between /h/ and the vowel were adjusted to align with clear appearance of an F1, and the boundaries between the vowel and the coda consonant were aligned to a simultaneous rapid cessation of most or all formants.

[6] The difficulty for some native talkers to produce [ʊ] when reading *hodd* might be due to frequency effects. Forms with stressed [ʊ] are few in the Swedish language, and phonotactically similar words are most often pronounced as [ɔ] (see e.g., Riad, 2014).

[7] English translations: "hod as in threat"(phonologically [uː]), "hodd as in cough"(phonologically [ʊ]).

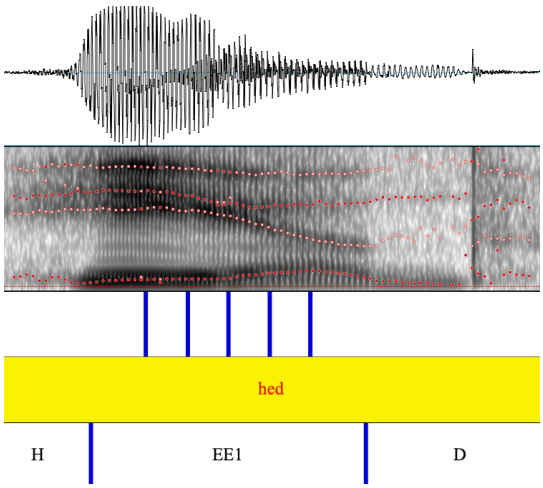

Figure 2. Example of Praat textgrid with annotated segment boundaries and measurement points for the automatic extraction of F1-F3 formant frequencies.

Extraction of phonetic cues

We used the Burg algorithm in Praat (Boersma and Weenink, 2022) to extract estimates of the first three formants (F1-F3) at five points of the vowel (20, 35, 50, 65, and 80 percent into the vowel). The following parameterization of the Burg algorithm was used:

- Time step (s): 0.01
- Max. number of formants: 5
- Formant ceiling (Hz): 5500
- Window length (s): 0.025
- Pre-emphasis from (Hz): 50

In addition to F1-F3, we automatically extracted vowel duration and the fundamental frequency (F0) across the entire vowel. The Praat scripts that extract this information are shared as part of the SwehVd OSF repository, allowing researchers to choose additional or alternative time points at which to extract formants.

In order to correct for measurement errors in the automatic extraction of cues, we estimated the joint multivariate distribution along all five extracted cues (F0, F1, F2, F3, and vowel duration) for each unique combination of vowel and talker. This approach allowed us to detect outliers relative to the joint distribution of the five cues for that vowel and talker. Points outside of the 0.5th to 99.5th quantile of the multivariate Gaussian distribution of each vowel were identified, checked for measurements errors, and corrected. For measurements of the first three formants, we first checked the segmentation boundaries in the Praat textgrid and then manually measured new formant values using visual approximation of time points and Praat's function *Formant: Formant listing* or manually reading off the spectrogram. Segmentation boundaries were also checked for the identified vowel duration outliers. For measurements of F0, we extracted new estimated F0s across the vowel, after changing the pitch range settings. Given that there were still instances of pitch halving after measurement correction, in order to be conservative, we also checked all F0 values below the point of intersection between the two halves. The database available via OSF reports

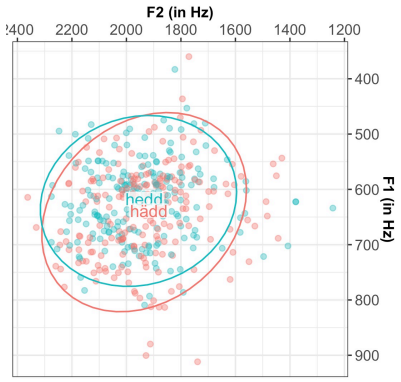

Figure 3. The *hedd* and *hädd* words in the SwehVd vowel data in unnormalized F1-F2 space. Points show recordings of the *hedd* and *hädd* words ([ɛ]) by the 23 female native talkers in the database, averaged across the five measurement points within each vowel segment. Word labels indicate word means across talkers. Since *hädd* and *hedd* resulted in the same allophone, we exclude *hädd* from this and all other visualizations below. This facilitates comparison of, for example, densities across vowels (see diagonal of Figure 5).

these corrected values (since the recordings and annotation grids are also available via OSF, other researchers can easily derive alternative measurements).[8]

The procedure of adding written guides to *hod* and *hodd* to facilitate vowel identification was mostly successful, however not for all talkers. Some talkers corrected themselves after one trial, others failed to produce the intended vowel altogether. The SwehVd database contains columns for both the targeted vowel category, and the vowel category that the talker actually produced (as annotated by the first author).

For the vast majority of talkers, *hädd* productions elicited the same vowel as *hedd* (see Figure 3). This confirms the common assumption that the short allophone to /e/ neutralizes with the short allophone to [ɛ] in Central Swedish.

Characterizing vowel productions in SwehVd

Figure 4 visualizes the vowel data from the SwehVd in F1-F2 space. The plot highlights the density of the Central Swedish vowel space, the categories are numerous and closely located. Category overlap is especially large among some of the high vowels (e.g., [iː] & [yː]; [uː], [oː] & [ʊ]). The contextually conditioned allophone [æ], almost completely overlaps with the long [ɛː], whereas the contextual allophones to [ø] are more separated. Not all contextual allophones are articulated lower (higher F1) in relation to their phonemes (compare e.g., Riad, 2014). They are, however, all articulated further back (lower F2).

In line with Riad (2014, cf. Table 2 above), the short vowels are overall more centralized and form a more condensed space, whereas the long vowels are more dispersed.[9] Differences in formant patterns between long and short vowels have been found to be smallest for the allophones to [ɛ] and [ø], and largest for /u/ ([ʉː] and [ɵ]), and /a/ ([ɑː] and [a]) (e.g., Fant, 2001; Kuronen, 2000). This

[8] We note that outlier detection and correction was based on raw, rather than transformed/normalized, cue values. For the studies we report below, this potentially introduces a bias *against* normalization. If anything, the present study is thus likely to under-estimate the effects of normalization.

[9] [ɪ] and [ʏ] are, however, more fronted than their long counterparts, which does not replicate previous descriptions of Central Swedish. We elaborate on this result in other ongoing work (Persson, 2023).

pattern does not entirely replicate here. We do see large differences in F1-F2 for the allophones to /u/ and /a/, but also substantial differences between [ɛ] and [ɛː]. Formant differences are apparent even for some category distinctions for which quantity has been found to be the primary cue (as shown in perceptual studies, see e.g., Behne et al., 1997; Hadding-Koch and Abramson, 1964), ([ɛː] - [ɛ], [øː] - [ø], [iː] - [ɪ], and [oː] - [ɔ]).

Figure 5 visualizes the vowel data from the SwehVd database for all pairwise combinations of five cues: F0, F1, F2, F3 and vowel duration. As is to be expected, vowels differing in quality are most separated in the F1-F2 plot, indicating the two cues most important for vowel category distinction. However, the F1-F3 and F3-F2 plots both display less overlap between the high vowels [iː], [yː] and [ʉː], comparing to when plotted along F1-F2. The increased separation of these categories along F3 in vowel production data could point to the importance of F3 for some category distinctions, as found in previous studies (see e.g., Fant et al., 1969; Fujimura, 1967; Kuronen, 2000, for [iː] and [yː] categorization).

Also as expected, duration is the primary cue that distinguishes vowel quantity: in the last column of Figure 5, the short vowels cluster on the left, and the long vowels on the right. They are separable, but overlapping. Overall, the short vowels seem to display less variability in duration than the long vowels, a common pattern for measures with a lower bound. In addition to duration, F1-F3 can also carry information about vowels differing in quantity. This is evident, for example, for [iː] vs. [ɪ], [yː] vs. [ʏ], [ʉː] vs. [ɵ], [ɑː] vs. [a], [ɛː] vs. [ɛ] in F1-F2 space, and for [iː] vs. [ɪ], [yː] vs. [ʏ], [ʉː] vs. [ɵ] in F2-F3 space.

Finally, the densities along the diagonal of Figure 5 suggest that F0 carries the least information about vowel identity, exhibiting the least between-category separation, followed by F3. This, too, is not surprising: while some accounts use F0 to *normalize* F1 and F2 (e.g., Miller, 1989; Syrdal and Gopal, 1986), F0 is not considered an important cue to vowel identity by itself (for demonstrations that F0 can, however, have strong *in*direct effects on vowel categorization, see Barreda and Nearey, 2012; Barreda, 2020).

STUDY 1: COMPARING THE EFFECTS OF NORMALIZATION ACCOUNTS ON BETWEEN- VS. WITHIN-CATEGORY VARIABILITY

In Study 1, we follow previous research on the effects of different normalization accounts on category variability, by evaluating how effectively different normalization accounts reduce the within-category variability of Central Swedish vowels. We visualize the vowel space under different normalization accounts, and assess the effects on vowel category variability by calculating a measure of category separability. To anticipate one take-home point of Study 1, the results highlight important shortcomings of separability indices in evaluating normalization accounts.

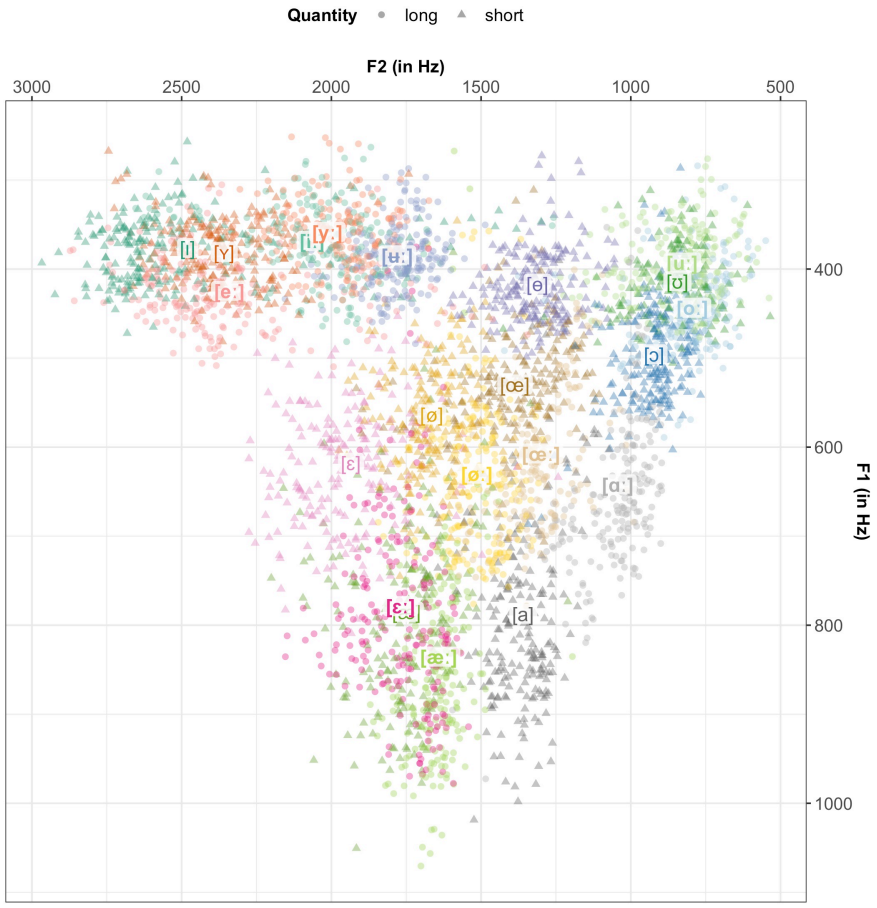

Figure 4. The SwehVd vowel data in unnormalized F1-F2 space. Points show recordings of each of the 21 Central Swedish vowels by the 23 female native talkers in the database, averaged across the five measurement points within each vowel segment. Vowel labels indicate category means across talkers. Long vowels are boldfaced. Vowels that mismatched intended label are excluded (1.23976% of all recordings). Note that the F1 and F2 axes are reversed. We follow this convention whenever plotting vowels in the F1-F2 space.

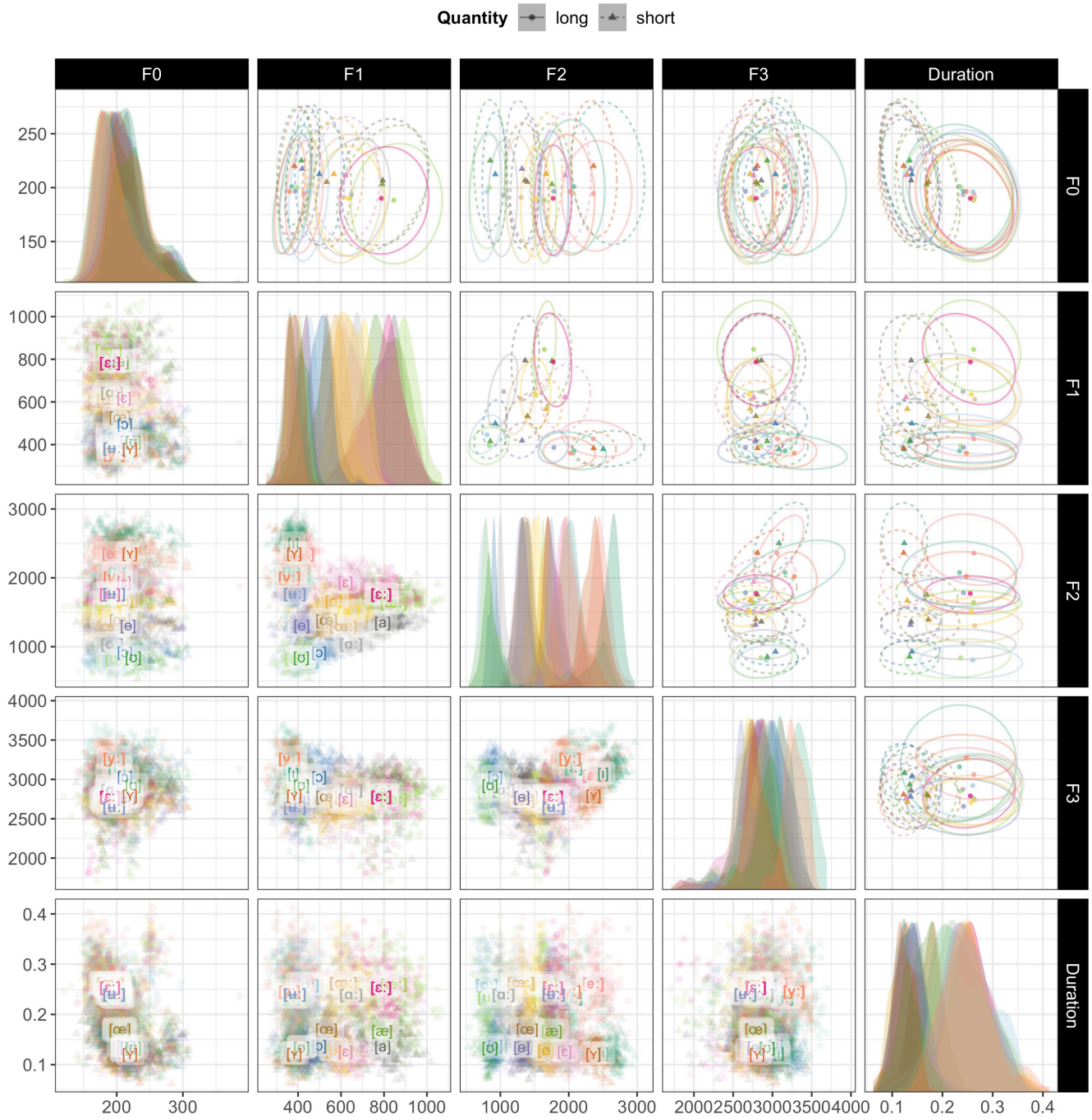

Figure 5. The same data as in Figure 4 but for all pairwise combinations of five cues: F0, F1, F2, F3, and vowel duration. The primary purpose of this figure is to provide an overview of the SwehVd data. Additionally, comparisons across the panels sheds light on which cues carry information about vowel quality and vowel quantity, respectively. Note that, unlike in Figure 4, axis directions are not reversed. **Panels on diagonal:** marginal cue densities of all five cues. **Lower off-diagonal panels:** each point corresponds to a recording, averaged across the five measurement points within each vowel segment. Vowel labels indicate category means across talkers. Long vowels are boldfaced. **Upper off-diagonal panels:** Same data as in the lower off-diagonal panels but showing bivariate Gaussian 95% probability mass ellipses around category means. This makes it more obvious, for example, that long and short vowels are primarily distinguished by vowel duration (top right panel).

Table 3. Previous studies comparing the effectiveness of normalization accounts in reducing within-category cue variability

Language investigated	Article	Speech materials	Normalization accounts	Approach	Best two performing
US English	Barreda & Nearey, 2018	120,000 simulated languages (of 5 or 9 vowels) modeled on Hillenbrand et al.'s (1995) data (98 female/male child/adult talkers * 12 vowels)	Nearey2, Lobanov, log-mean in linear regression framework	distance between means (Eucledian distance)	log-mean in linear regression framework (1), Nearey2 (2)
	Clopper, 2009	2 female/male talkers from Ohio (1 token * 10 vowels)	Bladon et al.'s scale factor of 1 Bark (1994), Syrdal & Gopal, Nordström & Lindblom, Nearey1, Nearey2, Watt & Fabricius, Gerstman, Lobanov, Miller	variance reduction (visual inspection)	Nearey, Watt & Fabricius, Gerstman, Lobanov (no order)
	Hindle, 1978	Peterson & Barney's (1952) database; 19 female/male talkers from Philadelphia + 60 telephone informants (minimum 3 tokens per category; analysis focus on /ay/)	Nearey2, Nordström-Lindblom, Sankoff-Shorrock-McKay	distance between means, variance reduction (regression)	Sankoff (1)
	Kohn & Farrington, 2012	Longitudinal data from 10 female/male African American talkers from North Carolina (approx. 10 tokens * 10 vowels * 5 ages)	Lobanov, Gerstman, Nearey1, Nordström & Lindblom, Syrdal & Gopal/Thomas, Watt & Fabricius	variance reduction (regression)	Lobanov (1), Gerstman, Watt & Fabricius (2)
	Labov, 2010	Peterson & Barney's (1952) database; Philadelphia/Linguistic Change and Variation project (120 female/male talkers, stratified for age, sociolinguistic factors)	Nearey2, Nordström-Lindblom, Sankoff-Shorrock-McKay	distance between means (F-statistics)	Sankoff (1), Nearey2 (2)
US English, Norwegian, Swedish, German, Danish, Dutch	Disner, 1980	Differing number of tokens, vowels, and phonetic contexts across the six languages	Gerstman, Lobanov, Nearey2, Harshman's PARAFAC model	variance reduction (visual inspection)	Nearey2 (1), Lobanov (2)
UK English	Fabricius, Watt & Johnson, 2009	20 old/young female/male talkers of Received pronunciation (11 vowels); 6 old/young female/male talkers of Aberdeen English (8 vowels in different phonetic contexts)	Watt & Fabricius, Lobanov, Nearey1	variance reduction (SCV in talker-means)	Lobanov (1), Watt & Fabricius (2)
	Flynn & Foulkes, 2011	20 old/young female/male Nottingham talkers (mean 180 recordings per talker; categories not reported)	log-transformation (base 10), log-transformation (natural), Mel, ERB, Bark (*2 gender-specific versions), Syrdal & Gopal, Nordström (*2 gender-specific versions), LCE, Gerstman, Lobanov, Watt & Fabricius (* 4 versions), lettER, Nearey (*4 versions)		Gerstman (1), LCE (2)
Russian	Lobanov, 1971	5 female/male talkers (9 vowels in different phonetic contexts)	linear compression or expansion (Fant, 1960), Gerstman, Lobanov	distance between means	Lobanov (1), Gerstman (2)

Table 3 lists previous studies that have compared normalization accounts in terms of their effectiveness in reducing inter-talker variability. These studies varied in the specific metrics they employed to assess the effects of normalization, the languages they studied, and the conclusions they arrived at. Two generalizations emerge from Table 3. First, transformation to a perceptual scale alone does not seem to be sufficient to reduce inter-talker variability (see also Adank et al., 2004; Carpenter and Govindarajan, 1993; Clopper, 2009; Escudero and Bion, 2007; Flynn and Foulkes, 2011; Kohn and Farrington, 2012). Second, normalization accounts that include centering and/or standardizing seem to perform best in reducing inter-talker variability (see e.g. Barreda and Nearey, 2018; Disner, 1980; Fabricius et al., 2009; Kohn and Farrington, 2012; Labov, 2010; Lobanov, 1971). When Lobanov and Gerstman normalization—both involving standardizing—were included in a study, they often rank among the top two performing accounts. Of note, Nearey normalization (Nearey, 1978) seems to perform well even though it does not involve the computationally more complex operation of standardizing. This suggests that simple centering of formants relative to the talker's mean *might* be sufficient to achieve significant variance reduction (but see Disner, 1980 for Swedish, which is revisited in this study).

Methods

Speech materials

We use the SwehVd database with some exclusions. Since we are interested in assessing the effects of normalization, we excluded any productions on which the talker did not produce the targeted vowel. We then excluded all talkers (N = 7) with fewer than 5 remaining recordings for at least one of the vowels. Five of these talkers rarely ever produced the targeted [ʊ] vowel despite our recording instructions (see Extraction of phonetic cues). Instead, they often mispronounced the vowel. This left data from 16 female native talkers, with on average 797 (se= 2.5) tokens per vowel (range = 765 to 815), for a total of 16730 observations.

Since our goal is to obtain a reliable estimate of the formant values during the steady state of the vowel, we use only the three formant measurements extracted from the middle of the vowel (at 35%, 50%, and 65% into the vowel).[10]

We also exclude all *hädd* productions, as they elicited the same vowel as *hedd* (see Extraction of phonetic cues). This way, we have about equally many tokens from all vowels, simplifying the cross-validation procedure presented below and facilitating visual comparisons across vowels in our figures.

Cues included in the normalization

We compare the effect of different normalization accounts on the variability of Central Swedish vowels under three different assumptions about the relevant cues. The first comparison follows most previous research and focuses on the two primary cues to vowel perception, F1 and F2. The second comparison considers F3 in addition to F1 and F2, following Adank et al. (2004), Barreda and Nearey (2018), Nearey (1989) and Syrdal (1985).[11] Finally, the third comparison includes F0 and duration in addition to F1-F3. Since Syrdal and Gopal (1986)'s bark-difference model only considers

[10] While this is the approach most commonly employed in the literature, it has the potential downside that co-articulation might affect formant values at the measurement points differently for long and short vowels (since the long and short vowels differ in overall duration). An alternative approach would be to extract formants at fixed durations (e.g., 30 ms) after the vowel onset and before the vowel offset. Since the findings we present below do not indicate any systematic differences in the performance of normalization accounts between long and short vowels, we do not consider this issue further here.

[11] Some of these studies additionally included F0 (Adank et al., 2004; Nearey, 1989; Syrdal, 1985). However, since F0 is a cue that can display substantial cross-talker variability without directly contributing much information to vowel categorization (recall Figure 5), its inclusion can reduce the informativeness of the separability index. We therefore decided to add only F3 to F1-F2 in the second evaluation.

normalization along two dimensions—height, implemented as F1-F0, and backness, implemented as F2-F1—this account will only be included in the first comparison. Furthermore, given that C-CuRE is the only account that applies to any type of cue, we will consider duration as centered to each talker's mean (for the C-CuRE accounts), or as raw input (in ms; for all other accounts). To our knowledge, duration has not been considered in previous studies on normalization, presumably because the present study is the first to evaluate normalization accounts against a vowel system with a systematic quantity distinction (long vs. short vowels). We evaluate the effect on variability both separately for long and short vowels, and on all 21 vowels together.

Separability index

Previous studies have used different measures to assess the relative success of a normalization procedure in reducing inter-talker variability (see Table 3 and Nearey, 1989 for an overview on classification accounts). This includes assessing the reduction in variance or distance between means by visual inspection (e.g., Clopper, 2009; Disner, 1980; Hindle, 1978), or by calculating the reduction in within-category variance across talkers (e.g., Disner, 1980; Fabricius et al., 2009; Flynn and Foulkes, 2011; Hindle, 1978), or comparing the degree of separation between category means for unnormalized and normalized data, i.e., an F-ratio (e.g., Labov, 2010).

We will assess how distinguishable vowels become under different normalization accounts by calculating a separability index, as described in Equation (1). Following some previous studies (e.g., Labov, 2010), this separability index is essentially an F statistics, where the F statistics is the ratio of the within- and between-category variances:

$$\begin{aligned} \textit{separability index} &= \frac{\textit{between category } MS}{\textit{within category } MS} \\ &= \frac{\sum_{c=1,\ldots,K}(N_c - 1)}{K-1} \frac{\sum_{c=1,\ldots,K}(\bar{x}_c - \bar{x})^2}{\sum_{c=1,\ldots,K}\sum_{i=1,\ldots,N_c}(x_{i,c} - \bar{x}_c)^2} \end{aligned} \tag{1}$$

where K is the number of categories, N_c is the number of observations for category c, $x_{i,c}$ is the cue vector (for all cues considered in the calculation of the separability index) for observation i of category c, $\bar{x}_c$ is the cue mean vector for category c, and $\bar{x}$ is the overall cue mean vector. We calculated this separability index separately for each combination of normalization account, cues, and training-test fold, as described next.

Guarding against over-fitting: cross-validation

As shown in Table 1, many of the normalization accounts involve parameters that are set based on the data (e.g., Gerstman, 1968; Lobanov, 1971; McMurray and Jongman, 2011; Miller, 1989; Nearey, 1978). This raises the question of how much these parameters can be affected by outliers, or other issues such as over-fitting to the sample. Unlike previous work, we thus use 5-fold cross-validation to obtain 5 separate estimates of the separability index for each combination of normalization procedure and cues. Specifically, we randomly split the data for each unique combination of talker and vowel into 5 even parts (folds). On each of the five folds, we then fit the normalization parameters based on four of the folds (the training data) and evaluated the effects of the normalization on the fifth fold (the test data). This resulted in five separability indices for each combination of normalization procedure and cues.

Results

Visualizing the distribution of vowel productions

Figures 6 and 7 visualize the Central Swedish vowels in the test data, after applying the 15 different scale-transformations and normalization accounts for a visual inspection. For this purpose, we focus on F1 and F2 only. The SI includes similar pairwise correlation plots of all cues for all different normalization accounts we compare (see SI, Correlation matrices).

Visual inspection suggests a few initial observations. The most striking difference is perhaps between intrinsic normalization accounts (Syrdal and Gopal, 1986; Miller, 1989) and all other approaches, though it is not immediately visually obvious which type of approach achieves better separability. Second, transforming the vowels to a different perceptual scale does not seem to affect the vowel distributions much, besides a minor decrease in category variance for some of the vowels. Some transformations bring the vowel categories closer together, towards the center of the vowel space, e.g., ERB and semitones. Third, centering formants by subtracting each talkers' mean (McMurray and Jongman, 2011; Nearey, 1978) reduces some of the category variance, and as a result, increases the category separability. Transforming the vowel data into different scales prior to centering also seems to further improve separability (compare e.g., C-CuRE (Hz) and C-CuRE (semitones)). Overall, the top two performing accounts across the long and short vowels appear to be Lobanov (1971) and Nearey (1978). However, even for the best performing normalization accounts, there is still considerable category overlap. This involves some of the high long vowels, and some of the mid-center short vowels. This highlights the need to more systematically quantify the effects of normalization, as we do next for category separability and then in more depth in Study 2.

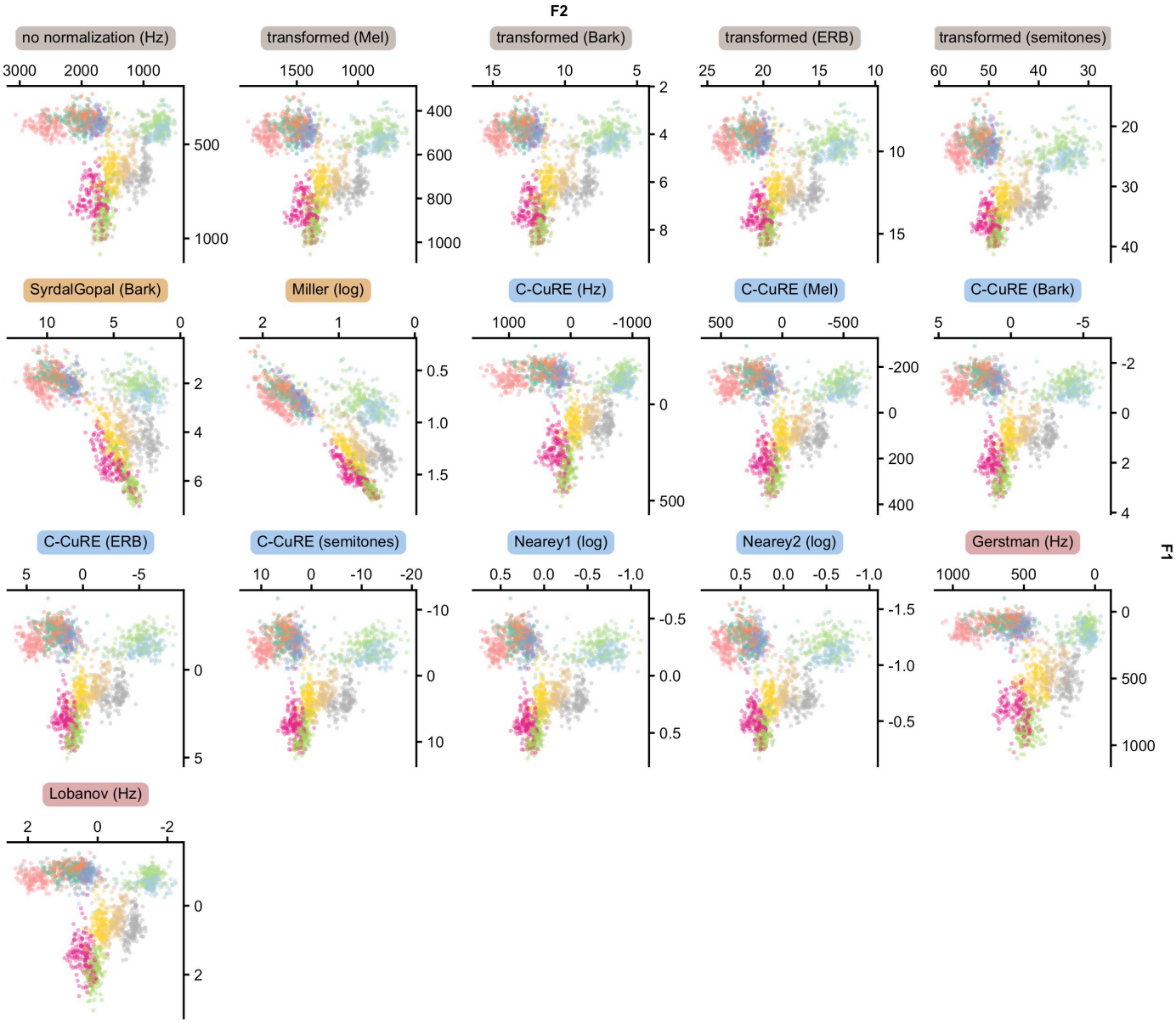

Figure 6. The 11 long vowels of Central Swedish when F1 and F2 are left unnormalized or transformed into a perceptual scales (**grey**), intrinsically normalized (**yellow**), or extrinsically normalized through centering (**blue**) or standardizing (**purple**). Each point corresponds to one recording, averaged across the five measurement points within each vowel segment. Each panel combines the data from all five test folds.

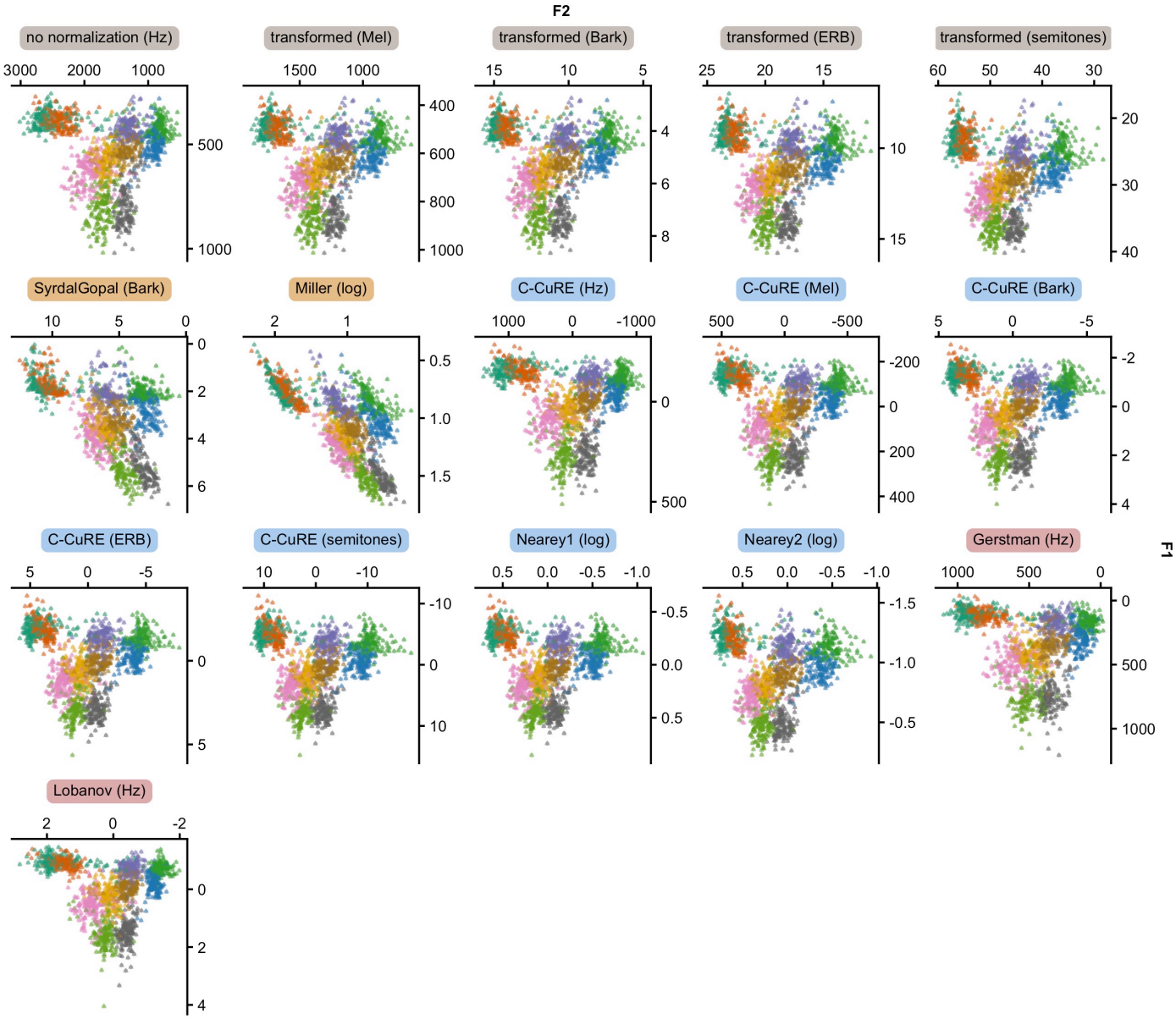

Figure 7. The 10 short vowels of Central Swedish when F1 and F2 are left unnormalized or transformed into a perceptual scales (**grey**), intrinsically normalized (**yellow**), or extrinsically normalized through centering (**blue**) or standardizing (**purple**). Each point corresponds to one recording, averaged across the five measurement points within each vowel segment. Each panel combines the data from all five test folds.

The effect of normalization accounts on the variability of Central Swedish vowels

In order to assess the effects of normalization accounts on category separability, we calculate a separability index under different assumptions about the relevant cues and the size of the vowel space (the long and short vowels separately, or the entire space). Before we evaluate how category separability is affected by normalization in F1-F2, F1-F3, and F0-F3 and duration space, we look at how the normalization accounts affect the separability of vowels along each cue separately (Figure 8). As we show below, this is helpful in understanding the subsequently presented results for combinations of cues.

For F1 (first row of Figure 8), we see a clear advantage for centering (in blue) and standardizing (in purple) compared to transformations (in grey) and intrinsic accounts (in yellow). In particular Lobanov normalization seems to maximize category separability along F1, at least for the long vowels and all vowels together. Notably, the accounts pattern differently along F2 (second row of Figure 8). Overall, differences between accounts are much smaller along F2, and the clear advantage of centering and standardizing accounts along F1 does not extend to F2.

An altogether different picture is observed for F3. Compared to F1 and F2, the intrinsic account (Miller) performs substantially better in separating categories along F3, while all other accounts perform poorly. This result is surprising: one of the downsides of intrinsic approaches that has been noted in previous work is their sensitivity to measurement error (Thomas and Kendall, 2007). This sensitivity is caused by the fact that intrinsic accounts use a single measurement for normalization, rather than the less noisy estimates resulting from aggregating across segments that are used in extrinsic accounts. Since the third formant is often described as more difficult to reliably estimate than other formants (leading to more measurement error), F3 would be expected to be particularly affected by this weakness of intrinsic accounts.

Yet, further visualization in Figure 9 confirms that F3 indeed separates categories particularly well when intrinsic normalization is applied. Compared to other accounts, Miller (1989) seems to be particularly successful in separating vowels that differ in lip rounding. For example, Miller (1989) separates two clusters among the high and mid-high vowels, one consisting of the back vowels [oː] and [uː], and the other one of the front [iː], and rounded [yː] and [ʉː]. One possible explanation for this result is that intrinsic normalization is indeed particularly effective for F3, and that our correction of measurement errors—equally applied to all formants—effectively reduced the issue with F3 measurement errors (presumably the human brain, too, can do better than an uncorrected Praat algorithm without error correction). As we show below, this result for F3 carries over to any combination of cues that includes F3. It is, however, an artifact of using category separability to assess the effectiveness of normalization, as we show in Study 2. We elaborate on this issue in the discussion of Study 1.

Returning to Figure 8, normalization does not increase category separability for F0. This is expected given that F0 is known to affect vowel separability primarily through its indirect influence on the interpretation of other formants (e.g., Barreda and Nearey, 2012; Barreda, 2020). Finally, for duration all of the C-CuRE accounts group together against the remaining accounts. This, too, is expected since all other accounts are formant-specific and thus do not normalize duration. In summary, the five cues contribute to category separability in different ways, and this is reflected in varying effectiveness of different normalization accounts. We also note that the best performing normalization account for any combination of cues and vowel qualities is typically never significantly better than the next best performing model (the 95% confidence intervals of the best model overlap

Figure 8. Separability indices by normalization accounts for long vowels, short vowels, and long and short vowels together (columns), shown for each of the five cues considered in this study (rows). Labels indicate mean across the five test folds. Intervals show average bootstrapped 95% confidence intervals across the test folds. Note that the ranges of the y-axes varies across plots.

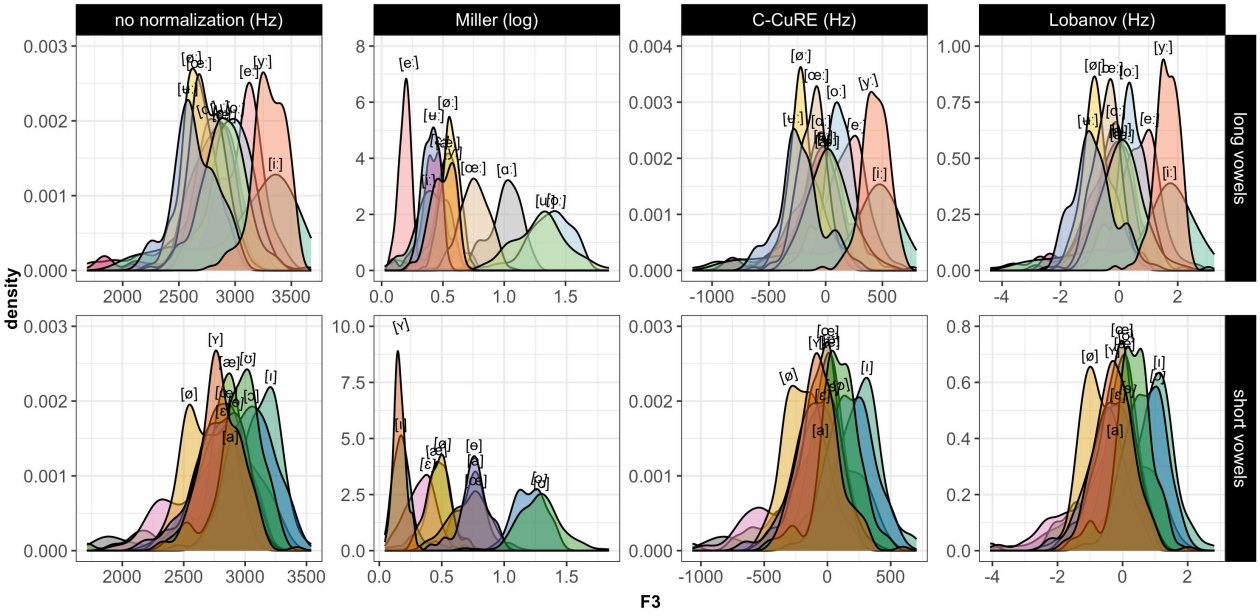

Figure 9. Category densities along F3 illustrates the effectiveness of vowel-intrinsic normalization for this cue. Here shown for Miller, compared to vowel-extrinsic accounts that center and/or standardize cues. For reference, densities in the absence of normalization are also shown.

with the mean of the next best model). In fact, for many combinations of cues and vowel qualities, many of the models perform similarly.

Next, we summarize how normalization affects category separability when combinations of the fives cues are considered. Figure 10 shows the separability index for the different normalization accounts for three different combinations of cues. For the first row of Figure 10, we followed most previous research in assessing category separability for the combination of F1 and F2 (e.g., Disner, 1980; Fabricius et al., 2009; Flynn and Foulkes, 2011; Hindle, 1978; Labov, 2010). Accounts that center against the talker's overall formant mean (in blue) are among the best performing normalization accounts. No matter the assumed perceptual scale, centering always improves category separability. Standardizing accounts (in purple), primarily Lobanov (1971), also perform well at separating categories, more so for the long vowels. However, scale transformations (in grey), and intrinsic accounts (in yellow), do not improve category separability compared to unnormalized Hz, at least not when assessed on the long vowels or the entire vowel space.

The remaining rows of Figure 10 compare normalization accounts when F3 (second row) or F0, F3, and duration are included (third row). Overall, the category separability is now lower, a result of how the accounts affect category separability along the cues added (see Figure 8). The most drastic change in performance concerns the intrinsic Miller (1989) and the standardizing accounts. When including F3, Miller (1989) performs as well or better, in absolute numbers, as when evaluated on only the combination of F1 and F2, thereby increasing its performance relative to other accounts. This increase in performance might be particularly pronounced for languages like Swedish, where F3

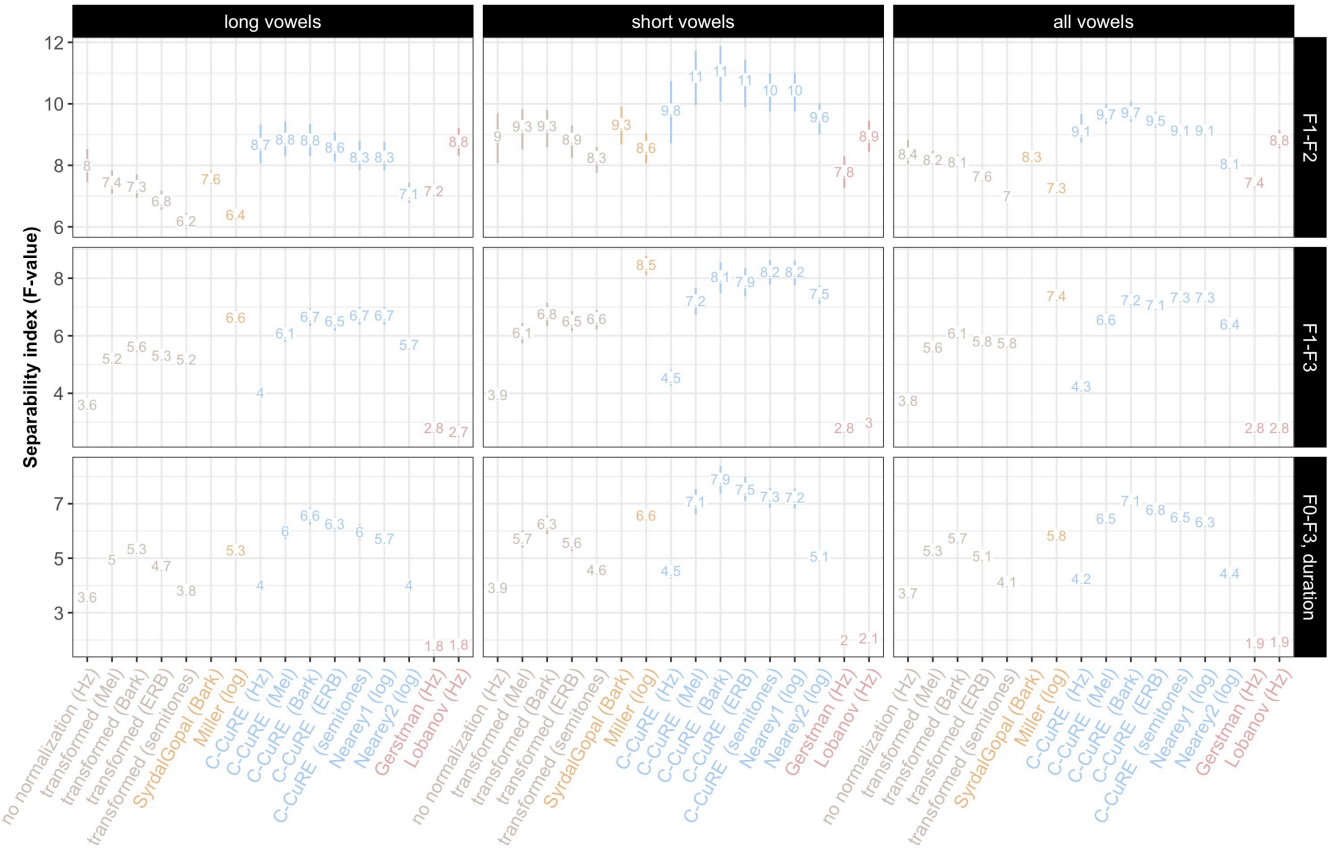

Figure 10. Separability indices by normalization accounts for long vowels, short vowels, and both long and short vowels together (columns) shown for three different combinations of cues (rows). Labels indicate mean across the five test folds. Intervals show average bootstrapped 95% confidence intervals across the test folds. Note that the ranges of the y-axes varies across plots.

carries important information about lip rounding and thus vowel identity. In contrast, performance of standardizing accounts drops substantially if F3 or any other cue besides F1 and F2 is included.[12] This mirrors what was found when assessing category separability separately for each cue (Figure 8).

Finally, looking across all three rows, category separability is consistently higher for short than long vowels. The same pattern is evident for each cue separately in Figure 8. This result might initially be puzzling, given that previous descriptions of Central Swedish vowel inventories characterize the inventory of short vowels as being more centralized and more densely clustered (e.g., Kuronen, 2000; Riad, 2014). Indeed, this claim seems to hold for SwehVd—compare Figures 6 and 7. However, the short vowels exhibit less category variability and less category overlap, making them overall more separable.

[12] We confirmed this by conducting additional comparisons using only F1, F2 and F0, or only F1, F2 and duration. For both of these comparisons too, we found that standardizing accounts perform poorly.

Discussion

When only F1 and F2 are considered, as in most previous work on vowel normalization, we find that extrinsic centering and standardizing accounts achieve the best category separability. Within these two types of accounts, there is considerable variability. For example, among the intrinsic accounts, Miller performs worse than Syrdal & Gopal, among the extrinsic accounts, versions of C-CuRE seem to consistently perform best. It is also worth noting, however, that there is never a single account that performs significantly better than all other normalizations. This points to the inherent similarities across normalization accounts, and perhaps limitations of the approach taken here (and in some previous work). We return to this point in the general discussion. Regardless of these caveats, the findings for F1 and F2 of Study I, revise the results of Disner (1980) for Swedish, and instead replicates previous findings for the other Germanic languages in Disner's sample as well as the majority of previous studies on other languages (e.g., Fabricius et al., 2009; Flynn and Foulkes, 2011; Labov, 2010).

However, when F3 is considered along with F1 and F2, this result does no longer hold. Key to understanding this result and what it says about the suitability of category separability as a measure of normalization accounts is Figure 8: while extrinsic normalization performs better than other approaches for F1 and F2, the absolute differences in performance are small compared to the advantage of the intrinsic account observed for F3. Combined with a seemingly innocuous aspect of the separability index in Equation (1), this allows separability along F3 to dominate separability along the other cue dimensions. Our separability index takes the *sum* of (squared) distances along each cue dimension, essentially assuming that the effect of all cues is simply a sum of each cue's effect considered separately. This means that the separability index cannot capture the *joint* effect of cues—whether, for example, one cue effectively separates one set of categories and another cue separates another set of categories, rather than both cues separating the same categories. The separability index thus cannot recognize, for example, that F1 and F2 capture largely complementary aspects of the vowel inventory (as evident in, for example, Figures 6 and 7).

This is not the only deficiency of the separability index or similar measures of category variability. The use of *squared* distances means that even a small number of observations located far away from the category mean can disproportionally affect the index. Consider, for example, the F3 densities in Figure 9. For non-intrinsic normalizations, some categories have low but non-zero densities far away from the mode. Because of the use of squared distances, this results in low category separability for these normalization accounts despite the fact that observations with such cue values are rare and thus not expected to have a large effect on the *average* perceptual separability of vowels. For the same reason (the use of squared distances), category separability can be high even if a cue separates only a small subset of categories (as is the case for F3), compared to another cue that more gradiently separates *all* categories (as is the case for F1 and F2; see Figure 5).

Finally, measures like the separability index suffer from a conceptual issue: the goal of speech perception is presumably not to reduce cue variability around the category mean but rather to increase the probability of correct recognition (of both linguistic and social information, where we focus on the former here). These two goals are not the same (see also discussion in Barreda, 2020). In sum, indices of variability and category separability like that in Equation (1) fail to adequately assess the expected consequences of normalization for perception, which is the primary interest of this study. Study 2 addresses this issue.

STUDY 2: COMPARING THE EXPECTED EFFECTS OF NORMALIZATION ON PERCEPTION

Study 2 takes a more direct approach to evaluating the expected effects of normalization for the perception of vowels. Specifically, we use a general model of speech perception, Bayesian ideal observers (e.g., Clayards et al., 2008; Nearey and Hogan, 1986; Norris and McQueen, 2008), to predict the vowel identities in the SwehVd database under different normalization accounts. We then compare normalization accounts based on the recognition accuracy that they achieve when the (un)normalized cues are fed into the otherwise identical categorization model. We evaluate the same normalization accounts as those investigated in Study 1. And, as in Study 1, we repeat this comparisons for different combinations of cues, and while categorizing just long vowels, just short vowels, or both long and short vowels together.

[13] Barreda & Nearey (2018) identify a mistake in the implementation of the Nearey2 account in Adank et al. (2004), so that the relative performance of Nearey2 reported by Adank and colleagues should be interpreted with caution.

Table 4. Previous studies comparing normalization accounts in terms of their predicted consequences for perception

Language(s) investigated	Article	Speech materials	Normalization accounts	Approach	Accuracy assessed	Best two performing
US English	Barreda, 2021	synthesized stimuli representing 6 talker types (based on data from 30 female/male talkers of California English (15 tokens * 11 vowels))	Nearey2, Watt & Fabricius, Lobanov	regression	against perceived category	Nearey2 (1), Watt & Fabricius (2)
	Carpenter & Govindarajan, 1993	Peterson & Barney's (1952) database, 75 female/male child/adult talkers (2 tokens * 10 vowels)	Bark, Mel, ERB, 2 log-transformations, Syrdal & Gopal, Miller, Nearey1, Nearey2, Gerstman, linear transformation (Watrous, 1993)	fuzzy ARTMAP, K-nearest neighbour	against intended category	linear transformation (1), Nearey1 (2)
	Cole, Linebaugh, Munson & McMurray, 2010	10 female/male talkers (3 tokens * 2 target vowels * 4 context vowels * 6 consonants)	C-CuRE	regression		C-CuRE (1)
	Johnson & Sjerps, 2021	Peterson & Barney's (1952) database, 75 female/male child/adult talkers (2 tokens * 10 vowels); Hillenbrand et al.'s (1995) database, 138 female/male child/adult talkers (1-3 tokens * 12 vowels)	Mean λ, F3 anchor, F1 anchor, Mean F* anchor (Sussman, 1986), Nordström, VTLN (Lammert & Narayanan, 2015), Nearey2, Gerstman, VTLN (ΔF), Nearey1, Watt & Fabricius, Lobanov, Miller, Syrdal & Gopal	support vector machine classification models		Lobanov (1), Watt & Fabricius (2)
	McMurray, Cole & Munson, 2011	Cole et al. (2010) database, 10 female/male talkers (1 token * 2 target vowels * 4 context vowels * 6 consonants)	C-CuRE	regression	against perceived category	C-CuRE (1)
	McMurray & Jongman, 2016	Jongman et al. (2000) database, 10 female/male talkers (1 token * 4 vowels * 8 fricatives)				
	Nearey, 1989	synthesized stimuli of male child/adult talker (based on male talker data from Fant, 1973, and Peterson & Barney, 1952)	intrinsic normalization, extrinsic normalization	response patterns (F-ratio)		extrinsic effects (1), intrinsic effects (2)
	Richter, Feldman, Salgado & Jansen, 2017	models based on Clopper & Pisoni's (2006) NSP vowel corpus, 60 female/male talkers, 6 varieties (5 tokens * 10 vowels); perceptual data from Feldman et al., 2009 (synthesized stimuli of male talker)	Vocal Tract Length Normalization (VTLN), Lobanov	discrimination model likelihoods		VTLN (1), Lobanov (2)
	Syrdal, 1985	Peterson & Barney's (1952) database, 75 female/male child/adult talkers (2 tokens * 10 vowels)	log-transformation, Bark, Syrdal's bark-difference model, Miller (2 accounts), Nearey1, Nearey2, Gerstman	linear discriminant analysis	against intended category	Nearey1 (1), Nearey2 (2)
Brazilian Portuguese & US English	Escudero & Hoffman Bion, 2007	models trained on 400,000 F1-F2 combinations generated on recordings of 8 female/male talkers (20 tokens * 7 vowels and 15 tokens * 11 vowels)	Nearey1, Lobanov, Gerstman	constraint rankings		Lobanov (1), Nearey1 (2)
Dutch	Adank, Smits & van Hout, 2004	160 female/male talkers, 8 varieties (2 tokens * 9 vowels)	log-transformation, Bark, Mel, ERB, Syrdal & Gopal, Lobanov, Nearey1, Nearey2[13], Gerstman, Nordström, Miller	linear discriminant analysis		

Table 4 lists previous studies that have compared normalization accounts in terms of their expected consequences for perception. While some previous studies have evaluated normalization accounts against listeners' responses in perception experiments (Barreda, 2021; McMurray and Jongman, 2016; Nearey, 1989; Richter et al., 2017), the majority of previous work has evaluated accounts against the category intended by the talker—i.e., against production data. Here, we follow this majority (though we plan to conduct perception studies on Central Swedish in the future).

Across languages and approaches, the studies in Table 4 yield similar results as those summarized in Table 3 for category separability. Normalization improves recognition accuracy compared to unnormalized cues. Perceptual transformations perform worse than intrinsic or extrinsic normalization at predicting perception from production data. Furthermore, the same normalization accounts that have been found to be most successful at reducing inter-talker variability (e.g., Lobanov, 1971; Nearey, 1978) are also often found to achieve higher recognition accuracy (e.g., Adank et al., 2004; Escudero and Bion, 2007; Johnson and Sjerps, 2021; Syrdal, 1985). Despite these overall similarities in results, it is important to keep in mind that the two approaches—category separability and models of perception—do not *have* to yield the same results. In the present study, we go beyond previous work by modeling perception of both vowel quality and vowel quantity over a particularly dense vowel space, and by considering additional cues.

As also shown in Table 4, previous work has employed a number of model types to compare the expected effects of normalization on perception, ranging from models based on phonological theory (e.g., optimality theory, Escudero and Bion, 2007), to more general models of categorization (e.g., linear discriminant analysis, Adank et al., 2004; Syrdal, 1985; k-nearest neighbors as in exemplar theory or ARTMAP, Carpenter and Govindarajan, 1993; Bayesian inference, Kleinschmidt et al., 2018; Richter et al., 2017; support vector machine classification models, Johnson and Sjerps, 2021), to general frameworks for data analysis (e.g., regression, Cole et al., 2010; McMurray and Jongman, 2016). We use ideal observers, rather than other approaches, because *all* of their degrees of freedom can be estimated from SwehVd. In contrast, k-nearest neighbor categorization requires the choice of a similarity metric, which can introduce one or more degrees of freedom into the modeling, and requires a choice for k. Similarly, linear discriminant analysis or regression introduce at least one degree of freedom for each cue considered. This means that any comparison of normalization accounts needs to be conducted over the entire range of possible values for these degrees of freedom, making comparisons computationally more demanding and interpretation of the results more difficult.[14] Bayesian ideal observers avoid this issue because of their assumption that listeners use and integrate cues *optimally*. As a consequence, the predicted posterior probabilities of all categories for a given acoustic input are determined by the combination of (1) the category-specific distribution of cues in the previous input and (2) the cue values of the input. Next, we describe ideal observers in more detail. After introducing ideal observers, we detail how this approach avoids the pitfalls of the separability index employed in Study 1.

Methods

Speech materials

Study 2 employs the same speech materials as in Study 1.

[14] The consequences of additional degrees of freedom that we describe here hold for comparisons of normalization accounts against *production* data (as done here and in the majority of research on normalization).

Ideal Observers to predict the consequences of normalization for perception

Ideal observers provide an analytical framework for estimating how a rational listener would optimally behave in response to input (here: n-way alternative forced-choice categorization). Ideal observer models have been found to provide a good qualitative and quantitative fit against human speech perception (e.g., Clayards et al., 2008; Feldman et al., 2009; Kleinschmidt and Jaeger, 2015; Kronrod et al., 2016; Norris and McQueen, 2008; Xie et al., 2021). Unlike most other models of speech perception, ideal observers in their simplest form—as employed here—have zero degrees of freedom in the link from production to perception: once the ideal observer is trained on phonetic data from a database of *productions*, its predictions about *perception* are not mediated by additional parameters (unlike, e.g., exemplar models, connectionist accounts, or neural networks). This removes researchers' degrees of freedom in the evaluation of normalization accounts, which is the reason we chose ideal observers for Study 2. We emphasize, however, that other researchers can download the R markdown document for this article (which contains the R code for our models) from OSF and substitute any other perceptual model for the ideal observers to assess the extent to which our choice of computational framework affects the findings reported below.

In line with influential theories of speech perception (e.g., exemplar theory, Johnson, 1997; Bayesian accounts, Luce and Pisoni, 1998; Nearey, 1990; Norris and McQueen, 2008; interactive-activation accounts and their offsprings, Magnuson et al., 2020; McClelland and Elman, 1986), ideal observers describe the posterior probability of a category as dependent both on the prior probability of the category in the current context, $p(category)$, and the likelihood of the acoustic input under the hypothesis that it originates from the category, $p(cues|category)$:

$$p(category|cues) = \frac{p(cues|category) \times p(category)}{\sum_c p(cues|category_c) \times p(category_c)} \tag{2}$$

The category prior, $p(category)$, describes how much the surrounding context favors each category. For Study 2, the choice of category prior cannot affect the qualitative results since category priors are independent of the cues and held identical across all normalization accounts.[15] We arbitrarily assume uniform category priors. Specifically, for ideal observers trained and tested on the long and short vowels separately, we model categorization as an 11- and 10-alternatives-forced-choice task, respectively, resulting in $p(category) = .091$ for the former and $p(category) = .1$ for the latter. For ideal observers trained and tested on the entire vowel space, we model categorization as a 21-alternatives-forced-choice task, resulting in $p(category) = .048$.

The likelihood, $p(cues|category)$, describes the distribution of cues for each category. Here, we follow previous work and assume multivariate Gaussian distributions to describe the cue likelihood (e.g., Clayards et al., 2008; Kleinschmidt and Jaeger, 2015; Kronrod et al., 2016; Xie et al., 2021). That is, we use the model in equation (3), where μ and Σ refer to the category mean and variance-covariance matrix of the category's multivariate normal distribution. In terms of representational complexity, the assumption of multivariate Gaussian categories strikes a compromise between exemplar storage (less representationally parsimonious, Johnson, 1997; Pierrehumbert, 2001) and cue integration over multiple separate univariate Gaussians (more parsimonious, Toscano and McMurray, 2010).[16]

[15] Specifically, category priors have a constant additive effect on the posterior log-odds of categories.

[16] Human perception is affected by an additional source of uncertainty beyond category variability: perceptual noise (for review, see Feldman et al., 2009). There is now evidence that including such perceptual noise in models provides a better fit against human behavior

$$p(category|cues) = \frac{\mathcal{N}(cues|\mu, \Sigma) \times p(category)}{\sum_c \mathcal{N}(cues|\mu_c, \Sigma_c) \times p(category_c)} \quad (3)$$

Paralleling Study 1, we trained and tested separate ideal observers for each combination of normalization account, cues, and training-test fold. Specifically, we use the exact same cross-validation folds as in Study 1. Each ideal observer was trained on the training portion of the folded unnormalized and normalized data, and subsequently evaluated on the held-out test fold. This means that the parameters of each normalization account (e.g., the cue means in C-CuRE) and the resulting category parameters (the μ_cs and Σ_cs for all categories) were set on the training data, and not changed for the test data. This reflects the realities of speech perception: although this is often ignored in evaluations of normalization accounts (e.g., Barreda, 2021; McMurray and Jongman, 2011), listeners do not *a priori* know the cue means, cue variance, etc. of an unfamiliar talker. Rather, listeners need to incrementally *infer* those statistical properties from the talker's speech input (for discussion and a model, see Xie et al., 2022). An additional advantage of cross-validation is that it gives us an estimate of the uncertainty about the model predictions. The performance of each ideal observer during test is assessed by calculating the ideal observer's predicted posterior probability of the *intended* category for each test token. This is identical to assuming Luce's choice rule for categorization (Luce, 1959), as in most models of speech perception and spoken word recognition.

Advantages compared to separability index and similar measures of category variability

By using a categorization model to estimate the consequence of normalization for perception, we avoid the pitfalls of the separability index discussed in Study 1. First, by using the density of an acoustic input under the multivariate distribution of all cues, the ideal observers we employ capture the *joint* effect of all cues. This captures that an input can be an improbable instance of a category based on one of its cue values but a probable instance given the values for the other cues. In particular, since we assume multivariate likelihoods, the model can account for category-specific covariances between cues. In the presence of strong correlations between cues, an acoustic input can be an improbable instance of a category based on the marginal distribution of each cue—i.e., if each cue is considered separately—but be a highly probable instance of that category if considered relative to the joint distribution of all cues (and vice versa). This is illustrated by Figure 11A.

Second, by normalizing the support for a category by the support for all other categories (the denominator in Equations (2) and (3)), ideal observers consider the perceptual consequences of an acoustic input *relative to all possible categories*. This means that a token that is relatively far away from its category mean does not necessarily result in low recognition accuracy. Rather, low recognition accuracy is only predicted if the relative position of the acoustic input in the acoustic-phonetic space makes it more probable that the input originated from another (unintended) category. This is illustrated in Figure 11B, and parallels human perception: e.g., while a more mid-fronted [ɑː] with high F1- and F2-values is atypical, human listeners are more likely to recognize it as a [ɑː] compared to a more high-back articulated, but equally atypical, [ɑː], presumably because the observed phonetics would be equally likely to occur if the talker intended a [oː]. Measures of

(e.g., Kronrod et al., 2016; Tan and Jaeger, 2022). Since Study 2 assesses the *relative* recognition accuracy of different normalization accounts, it is not immediately obvious how the inclusion of noise could affect our results. To avoid additional researchers degrees of freedom—such as the decision as to which acoustic or perceptual space (Hz, Mel, Bark, etc.) perceptual noise is additive in—we do not model the perceptual consequences of noise. To the best of our knowledge, no previous comparisons of normalization accounts has considered perceptual noise.

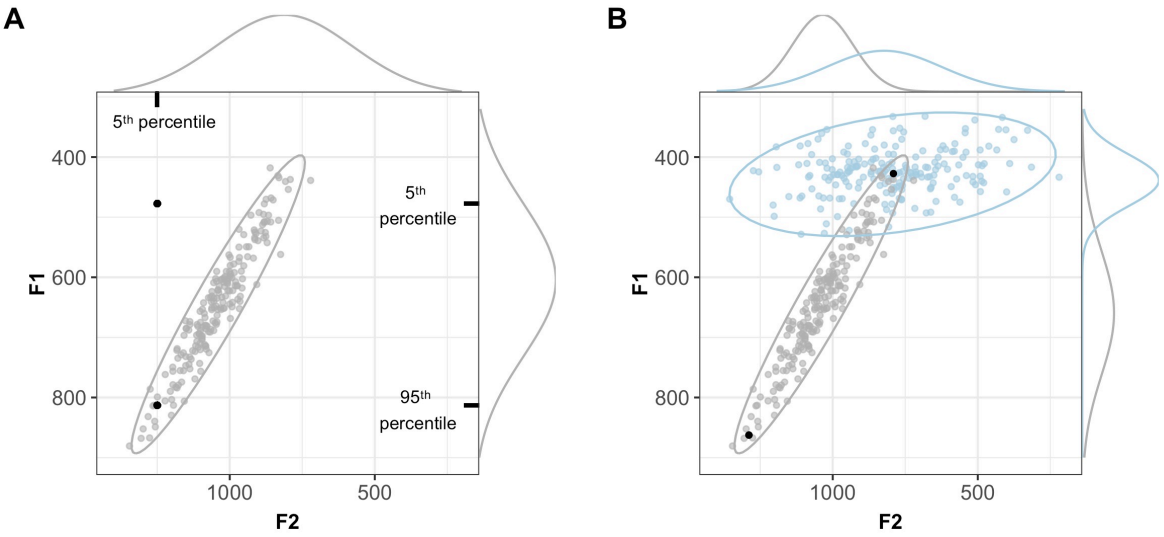

Figure 11. Using a perceptual model to evaluate normalization accounts avoids the pitfalls of separability/variability indices. **Panel A:** an acoustic token can be an improbable instance of a category if each cue is considered separately (the marginal densities along the sides of the plot), but highly probable if considered relative to the joint distribution of cues (the bivariate distribution indicated by the ellipse). **Panel B:** two acoustic tokens that are equally far from the mean of one category can have radically different consequences for perception, depending on where the tokens fall relative to other categories. Under the hypothesis that the two black points are instances of the gray category, they would be attributed the same separability index but radically different probabilities given the joint distribution of cues relative to the other category in the space. Both points are on the 90% highest density interval isoline.

between- vs. within-category variability like the separability index in Study 1, however, have no means of directly capturing this.

Finally, when an acoustic input is indeed improbable under its intended category relative to other unintended categories, the ideal observer model will correctly predict low recognition accuracy. Unlike the separability index in Study 1, however, the ideal observer will not disproportionally weight this single observation: it will simply be one of many observations—much like human speech perception does not collapse simply because one word (or one vowel) has been misunderstood. As we show next, these differences between the approaches taken in Studies 1 and 2 can, and sometimes do, cause differences in results.

Results aggregating across vowels

Figure 12 visualizes the unnormalized and normalized models' predictions for perception of Central Swedish vowels, under different assumptions about the relevant cues. As in Study 1, this figure aggregates results across vowels of a given type (long, short, both). We first discuss these results, and then briefly summarize how different vowels are affected by normalization.

The first observation is that—unlike for the separability index in Study 1—the relative performance of the different normalization accounts within each panel is remarkably constant across all panels (compare to Figure 10). Regardless of the combination of cues or the vowel types considered (long, short, both), transformation into a perceptual space does little to improve recognition accuracy, compared to unnormalized cues. Intrinsic normalization, too, does not improve recognition accuracy. And, unlike in Study 1, this result holds even when F3 is included as a

cue. This replicates previous work on Dutch (Adank et al., 2004) but conflicts with some evaluations of English (e.g., Syrdal, 1985). Adank et al. (2004) discussed whether the discrepancy in results might be attributed to implementations of the Bark-transformation, or to what Syrdal (1985) describes as language-specificity of the second dimension of Syrdal and Gopal (1986) normalization. The present results would seem to confirm this vulnerability of intrinsic normalizations. Extrinsic normalization, however, tends to substantially improve recognition accuracy (with the exception of Gerstman normalization). Depending on the specific combination of cues and the vowel qualities considered, the best-performing normalization model increases recognition accuracy by at least 46.8% (from 40.8% for unnormalized cues for all vowels when only F1-F2 are considered) to 81.2% (from 75.6% for short vowels when all cues are considered). The benefit of extrinsic normalization models, as well as the lower performance of perceptual transformations, replicates previous findings on other languages (e.g., Adank et al., 2004; Escudero and Bion, 2007; Nearey, 1989 found effects of both intrinsic and extrinsic accounts, but larger effects for extrinsic).

We also see that all models—even for unnormalized cues—perform substantially above chance. When long and short vowels are considered separately, the best ideal observers achieve recognition accuracies of 73.9% for long vowels and 81.2% for short vowels. For reference, in a recent perception experiment we conducted on the eight monophthongs of US English, L1-US English listeners achieved 71.1% accuracy in categorizing isolated hVd words (chance = 12.5%, Persson and Jaeger, 2023). The ideal observers for the Central Swedish vowel system thus achieve performance that is comparable to that of human listeners, at least when cues are normalized.

Looking across columns of Figure 12, short vowels are always recognized with higher accuracy compared to long vowels. This increase in performance cannot be explained by the small increase in the chance baseline alone (10% for the 10 short vowels, compared to 9.1% for the 11 long vowels). It conceptually replicates an initially surprising result of Study 1: while short vowels are more densely clustered in the center of the vowel space, and thus occupy a smaller perceptual space, they also exhibit less variability. Overall, this makes those vowels *easier* to recognize.

When long and short vowels are categorized together, performance of the ideal observers is comparatively poor unless vowel duration is included as a cue. This is expected given that vowel duration is the primary cue to vowel quantity. Of interest, however, is that even the inclusion of only F3 (second row) yields a substantial improvement in recognition accuracy, in line with Johnson and Sjerps (2021). Remarkably, once vowel duration is included, the best-performing ideal observer achieves 75% recognition accuracy across the 21 long and short vowels (compared to chance = 4.8%).

Looking across rows, we note that Lobanov normalization clearly performs best when only the first two formants are considered. Indeed, when only F1 and F2 are considered and all vowels are categorized (top right panel of Figure 12), Lobanov performs *significantly* better than any of the other normalization models (the 95% confidence intervals of Lobanov normalization does not include the mean of the second best performing model). However, this advantage of Lobanov normalization decreases and is no longer significant when additional cues are considered.[17]

[17] Indeed, when all five cues are considered for the categorization of all 21 short and long vowels, the best centering account perform numerically better (74%) than Lobanov normalization (72.7%). This is, however, an artifact of our decision to only center vowel duration—the primary cue to vowel quantity—for the C-CuRE model (paralleling Study 1). Separate modeling not shown here confirmed that Lobanov normalizaton achieves the same recognition accuracy as the C-CuRE models when duration is centered and combined with Lobanov-normalized formants (74.9%, 95%-CI: 73.9-75.9%).

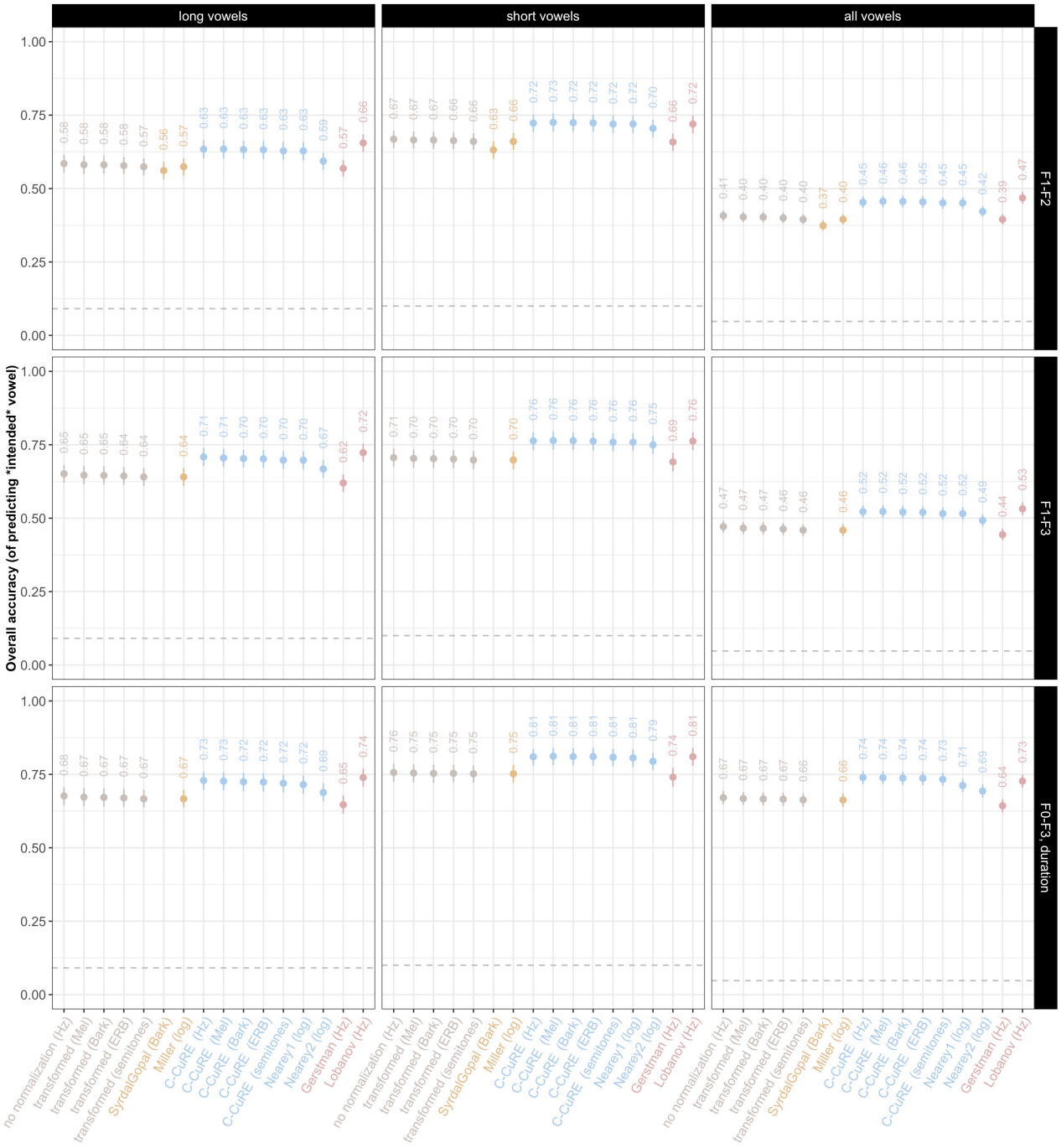

Figure 12. Predicted recognition accuracy of ideal observer under different normalization accounts for long vowels, short vowels, and long and short vowels together (columns), shown for for three different combinations of cues (rows). Labels indicate mean across the five test folds. Intervals show average bootstrapped 95% confidence intervals across the test folds. The dashed horizontal line indicates chance (different across columns because of the different number of long and short vowels).

Results for specific vowels

Finally, the use of a perceptual model lets us assess vowel-specific effects on the predictions of the ideal observers. While this is not the main focus of this paper, it might be of interest to other researchers. In the SI, we therefore plot both the predicted categorization accuracy per vowel in the different evaluations (see SI, Per-vowel categorization accuracy), as well as confusion matrices of the best and the worst performing ideal observers (see SI, Confusion and difference matrices), to further investigate *how* normalization improves recognition accuracy. Here, we briefly mention some of the main findings that emerge from these additional visualizations. For reference, Figure 13 visualizes the predicted recognition accuracy for five vowels that illustrate some of the vowel-specific effects across different types of normalization.

Unsurprisingly, some vowels are recognized with much higher accuracy than others—at least when uniform category priors are assumed, as we did here. This is a direct consequence of the position of the vowel in the acoustic-phonetic space, relative to neighboring vowels: the more neighboring vowels overlap with each other, the lower the accuracy with which they are recognized. Which vowels will benefit from normalization will thus naturally vary between languages, reflecting the language-specific properties of the vowel space. For instance, [iː] is often described as more easily recognized in previous work on other languages. This contrasts with our findings for Central Swedish: here, [iː] is part of the dense clustering of vowels along the height dimension and so has many close competitors. This highlights that recognition accuracy is due to the position of a vowel *relative* to its competitors (e.g., Peterson and Barney, 1952; Kuhl, 1991; Polka and Bohn, 2003), rather than its *absolute location* in the vowel space (e.g., [iː] being a peripheral vowel).

Also of interest is that not all vowels exhibit the benefit of normalization. In general, across evaluations, it seems to be vowels that are already recognized with relatively high accuracy that particularly benefit from normalization, which does not replicate previous studies that have included per-vowel accuracies (e.g., Adank, 2003; Syrdal and Gopal, 1986). In fact, Adank (2003) reports larger improvements in error rates after normalization for vowels with higher error rates prior to normalization. Finally, while there are minor differences across vowels in the relative goodness of different normalizations, the models that perform overall best also perform best on each vowel (in line with Adank, 2003). This further demonstrates the plausibility of these normalization accounts for perception.

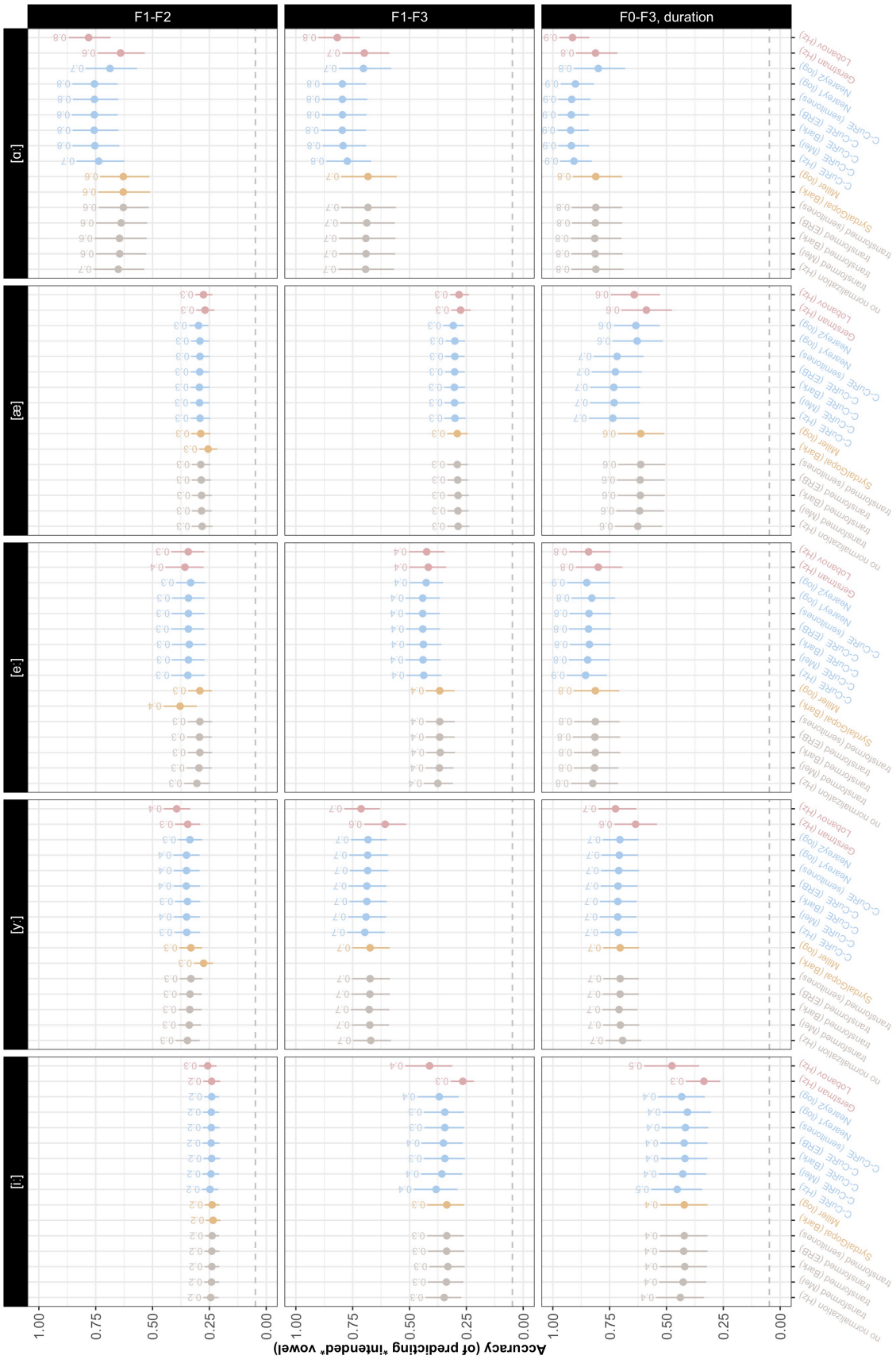

Figure 13. Predicted recognition accuracy of ideal observer under different normalization accounts for five of the 21 vowels. Point ranges indicate the average mean accuracy and average 95% bootstrapped CI across the five folds. Labels indicate mean across the five test folds. Chance level is indicated by grey line.

GENERAL DISCUSSION

We have compared low-level pre-linguistic normalization accounts against a new phonetically annotated database of Central Swedish vowels. We set out to evaluate how the different accounts affect the category separability (Study 1), and how they differ in predicted consequences for perception (Study 2). As we have already discussed the methodological shortcomings of separability and variability indices like that employed in Study 1, we focus here on the discussion of Study 2.

Previous work found that the types of normalization accounts that performed well on other languages did not seem to perform well on Swedish vowel data (Disner, 1980). However, as pointed out by Disner, the Swedish data differed from the data for other languages in that study, and the majority of studies on other languages. Here, we followed the majority of previous work on vowel productions and analyzed productions of hVd recordings. We find—both in Study 2 and in Study 1 (as long as additional cues are not included)—that the same accounts found in previous work to perform well on other languages also perform well for the dense vowel space of Swedish. Specifically, Lobanov and centering approaches—incl. Nearey normalization and C-CuRE normalization—were the top-performing accounts, replicating the pattern found in previous studies on other languages (e.g., Adank et al., 2004; Carpenter and Govindarajan, 1993; Escudero and Bion, 2007; Syrdal, 1985). This result suggests that the (somewhat) diverging results for Swedish in Disner (1980)'s study, were not caused by properties inherent to Swedish, but more likely were an artifact of the dataset employed by Disner. It also suggests that languages with dense vowel spaces do not necessarily require more complex normalization mechanisms.

Additionally, we find that the best-performing centering accounts (C-CuRE) often achieve performance that is statistically indistinguishable from the best-performing standardization accounts (Lobanov). This is the case, in particular, when all five cues were considered and all 21 vowels were included in the categorization (see footnote 17). Together with similar findings from research on consonants and supra-segmental categories (e.g., Apfelbaum et al., 2014; Crinnion et al., 2020; Kleinschmidt, 2019; Kulikov, 2022; McMurray and Jongman, 2011, 2016; Toscano and McMurray, 2015; Xie et al., 2021, 2022), this suggests that simple centering operations might be sufficient to maximize the benefits achievable by normalization. In the remainder, we first summarize some methodological considerations based on the present study, and then discuss limitations of our work, and how they can be addressed in future work.

Methodological considerations

We considered two methods of evaluations. In Study 1, we compared the reduction of within-category variability relative to between-category variability by calculating a separability index (i.e., an F-statistics). This methodology facilitated comparison to previous studies, that has most often assessed the efficacy of normalization accounts by means of increase in between-talker category overlap or category separability, or decrease in category variability (Disner, 1980; Fabricius et al., 2009; Flynn and Foulkes, 2011; Hindle, 1978; Labov, 2010). However, given the limitations inherent in this index, in the second study, we used Bayesian ideal observers to investigate how normalization accounts differ in predictions for perception. Previous studies have employed linear discriminant analysis (e.g., Adank et al., 2004; Syrdal, 1985), fuzzy ARTMAP or K-nearest neighbour classification (e.g., Carpenter and Govindarajan, 1993), or constraint ranking grammars (e.g., Escudero and Bion, 2007). While these different approaches will generally return similar results, we see two advantages with ideal observers. First, ideal observers remove researchers' degrees of freedom in the evaluation of normalization accounts (see Tan et al., 2021): the statistics

of the cue distributions in the training data fully determine the ideal observers' predictions. Second, representing the cue likelihood for each category as multivariate Gaussian distributions (e.g., Kronrod et al., 2016; Norris and McQueen, 2008; Xie et al., 2021), strikes a middle ground between less parsimonious models such as exemplar models (e.g., Johnson, 1997; Pierrehumbert, 2001), and more parsimonious models, such as models of cue integration over multiple separate univariate Gaussians (e.g., Toscano and McMurray, 2010).[18] However, like any other model of speech perception, the approach adopted here comes with a set of simplifying assumptions. All of the assumptions in the model, e.g., uniform priors, models of linguistic representations, the normalization accounts selected, can, however, be revisited and altered by the reader, as this paper is written in R markdown and all data and code is provided on OSF.

As a final note on methodology, we find the 5-fold cross-validation approach used for normalization of the acoustic data adopted here, advantageous for several reasons, the two most important being that: (1) it allows us to avoid over-fitting to the sample, while also (2) providing a more realistic reflection of how parameters used for normalization are incrementally inferred from the talker's speech input. Even though many of the commonly adopted normalization accounts involve parameters that are set based on the data, previous studies have rarely considered how these parameters might be affected by the specific dataset used for normalization (for exceptions, see e.g., Barreda and Nearey, 2018). In the present study, we have seen that accounting for researchers' uncertainty about the effects of normalization, highlights that many of the normalization accounts exhibit statistically indistinguishable performance—at least under the approach taken here and in the majority of previous work. This means that future studies should aim to increase statistical power, in order to determine the normalization mechanism that best describes human behavior. This will require even larger datasets, and/or by targeted sampling of vowel tokens for which predictions of different normalization accounts are maximally contrasted (see e.g., Barreda, 2021; and see Xie et al., 2022 for more general discussion of how to increase the statistical power to determine what mechanism underlie adaptive speech perception). Next, we close by discussing additional limitations of the present work and future directions.

Limitations and future directions

Four limitations of the present study, three of which are shared with most previous work, deserve discussion. First, the present study compared normalization accounts against speech from only female talkers of one regional variety of Central Swedish (Stockholm Swedish). In contrast, many previous studies included data from talkers of different genders (e.g., Barreda, 2021; Clopper, 2009; Cole et al., 2010; McMurray et al., 2011; McMurray and Jongman, 2016), and sometimes from talkers of different ages (e.g., Barreda and Nearey, 2018; Carpenter and Govindarajan, 1993; Flynn and Foulkes, 2011; Hindle, 1978; Johnson and Sjerps, 2021; Kohn and Farrington, 2012; Syrdal, 1985) and/or language backgrounds (e.g., Adank et al., 2004; Disner, 1980; Escudero and Bion, 2007; Fabricius et al., 2009; Labov, 2010; Richter et al., 2017). Given that age, gender, etc. tends to affect formants (and other cues) beyond talker-variability, it is likely that the inclusion of more diverse talkers would increase the lack of invariance problem. For example, we would expect the ideal observers over unnormalized cues (Study 2) to achieve lower recognition performance if vowel productions from male talkers would be included in the data. In short, the models in Study 2

[18] Parsimony here refers both to the number of degrees of freedom these models afford to the researcher *and* to the amount of information that listeners are assumed to store (for discussion, see Xie et al., 2022).

likely over-estimates the recognition accuracy that can be achieved for unnormalized cues if a more diverse range of talkers is considered.

What does that mean for the relative effect of normalization? To the extent that normalization successfully overcomes inter-talker variability that is caused by gender, age, and other social or physiological factors, we expect that the benefit of normalization accounts should show more clearly, relative to unnormalized cues. In this sense, the present study might *under*-estimate the relative benefits of normalization. Whether the *relative* performance of normalization accounts—i.e., the finding of primary interest to us—would differ if a more diverse range of talkers was considered is unclear. To the extent that vowel-specific accounts were originally developed specifically to eliminate physiological differences that are correlated with gender (as reviewed in, e.g., Johnson and Sjerps, 2021), it is theoretically possible that the high performance of general normalization accounts (e.g., C-CuRE, McMurray and Jongman, 2011) might not replicate when talkers of different genders are included. Future releases of the SwehVd database will contain data from male talkers, which will allow us or other researchers to revisit these questions.

Second, the present study aggregated acoustic-phonetic measurements taken at different points of the vowel (at 35%, 50%, and 65% into the vowel) into a single formant measurement. This follows previous comparisons of normalization accounts but is a simplifying assumption that should be revisited in future work. Formant dynamics carry important information for category distinctions (e.g., Assmann and Katz, 2005; Hillenbrand and Nearey, 1999; Nearey and Assmann, 1986), and are hypothesized to be of particular importance for some vowel distinctions in other varieties of Central Swedish (e.g., Kuronen, 2000). Prior to other consideration, this means that Study 2 likely under-estimates the recognition accuracy that could be achieved even from unnormalized cues alone. It is an open question whether the findings of primary interest—the relative performance of different normalization accounts—would be affected if formant dynamics were considered. Some normalization accounts, for example, consider normalization of such formant dynamics to take place *after* basic formant normalization (but before the mapping of cues to category representations, S. Barreda, personal communication, 01/06/2023). Future work could employ SwehVd to compare ideal observers or other classification models while taking into consideration formant measurements throughout the vowel.

Third, we only considered competing *normalization* accounts. This, too, follows previous research on normalization but is potentially problematic. As mentioned in the introduction, it is now believed that at least three different mechanisms contribute to adaptive speech perception, including not only normalization but also changes in category representations and decision-making (for review, see Xie et al., 2022). This has consequences for research on normalization. For example, Xie et al. (2021) compared normalization accounts against the production of prosodic phrasing in L1-US English, while also considering alternative hypotheses about listeners' ability to adapt category representations. Xie and colleagues found that the effectiveness of cue normalization is substantially reduced if listeners can learn and maintain talker- or group-specific category representations (as assumed in some influential theories of speech perception, exemplar models, e.g., Johnson, 1997; Pierrehumbert, 2001; Bayesian ideal adaptors, Kleinschmidt and Jaeger, 2015). Xie and colleagues only considered two general types of normalization, and their focus was on the interpretation of prosodic signals. But their results call for caution in interpreting studies like the present that do not consider the possibility of talker-specific representations—an assumption shared with basically all previous work on vowel normalization.

Similarly, as mentioned in the introduction, we limited our evaluation to a single level of normalization (and combinations of perceptual transformations and a single level of normalization). Some proposals, however, assume multiple separate normalization steps. For example, some accounts hold that evolutionarily early mechanisms first transform spectral percepts into a phonetic space (e.g., uniform scaling accounts, Barreda, 2020; Nearey, 1983), on which additional subsequent normalization might operate. There is also evidence that speech perception combines aspects of intrinsic and extrinsic normalization (Johnson and Sjerps, 2021 review relevant evidence from brain imaging; early behavioral evidence is found in Nearey, 1989). The present study—like most existing evaluations—did not consider these possibilities (for exceptions, see e.g., Barreda, 2021; Nearey and Assmann, 2007).

Fourth and finally, we followed the majority of previous work and evaluated normalization accounts against *production* data. This is potentially problematic, especially when measures like category separability or reduced cross-talker variability in category means are used to evaluated normalization accounts (as in our Study 1 and many previous studies). These evaluations essentially assume that the goal of speech perception is to make the perceptual realizations of the same category by different talkers as similar as possible in the normalized space (for an in-depth critique, see Barreda, 2021). However, the goal of speech perception is presumably to reliably infer the category intended by the talker,[19] and this aim does not necessarily entail perfect removal of cross-talker variability (as evidenced, for example, by the different findings of Studies 1 and 2).

To some extent, Study 2 addresses this potential issue by evaluating normalization accounts in terms of how well they predict the vowel category intended by the talker. However, if the goal is to explain human perception, the most informative evaluations of normalization accounts are arguably those that compare their predictions against *listeners'* behavior (for examples, see Barreda, 2020, 2021; McMurray and Jongman, 2016; Nearey, 1989; Richter et al., 2017; Xie et al., 2021). In short, approaches like that employed in Study 2 take an important step away from the most misleading evaluation of normalization accounts in terms of reduced category variability/increased category separability. Ultimately, however, normalization accounts should be evaluated in terms of how well they predict listeners' perception, not talker's intention.

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

AP proposed project idea, and both authors jointly developed the conceptual approach. AP designed SwehVd materials, recorded and annotated vowel productions, and coded cue extraction. AP coded data analyses and visualization with guidance from TFJ. Both authors contributed to the writing of the manuscript.

ACKNOWLEDGMENTS

We are grateful to OMITTED FOR REVIEW.

[19] Or somewhat more precisely, cooperative listeners aim to understand the meaning intended by the talker, and this inference is generally believed to benefit from the correct recognition of phonological categories, such as phonemes, syllables, or word forms. For discussion, see also Hume et al. (2016).

REFERENCES

Adank, P. (2003). *Vowel normalization: A perceptual-acoustic study of Dutch vowels*. Wageningen: Ponsen & Looijen.

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

Apfelbaum, K., Bullock-Rest, N., Rhone, A., Jongman, A., and McMurray, B. (2014). Contingent categorization in speech perception. *Language, Cognition and Neuroscience* 29, 1070. doi:10.1080/01690965.2013.824995.

Assmann, P. F., and Katz, G. S. (2005). Synthesis fidelity and time-varying spectral change in vowels. *Journal of the Acoustical Society of America* 117, 886–895.

Assmann, P. F., Nearey, T. M., and Hogan, J. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America* 71, 975–989.

Audacity, T. (2021). Audacity(R): Free Audio Editor and Recorder [Computer Application].

Barreda, S. (2021). Perceptual validation of vowel normalization methods for variationist research. *Language Variation and Change* 33, 27–53. doi:10.1017/S0954394521000016.

Barreda, S. (2020). Vowel normalization as perceptual constancy. *Language* 96, 224–254. doi:10.1353/lan.2020.0018.

Barreda, S., and Nearey, T. M. (2018). A regression approach to vowel normalization for missing and unbalanced data. *The Journal of the Acoustical Society of America* 144, 500–520. doi:10.1121/1.5047742.

Barreda, S., and Nearey, T. M. (2012). The direct and indirect roles of fundamental frequency in vowel perception. *The Journal of the Acoustical Society of America* 131, 466–477. doi:10.1121/1.3662068.

Behne, D. M., Czigler, P. E., and Sullivan, K. P. H. (1997). Swedish Quantity and Quality: A Traditional Issue Revisited. *Reports from the Department of Phonetics, Umeå University* 4, 81–83.

Bladon, A., Henton, C. G., and Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. *Language and Communication* 4, 59–69.

Boersma, P., and Weenink, D. (2022). Praat: Doing phonetics by computer [Computer program].

Bruce, G. (2009). "Components of a prosodic typology of Swedish intonation," in *Components of a prosodic typology of Swedish intonation* (De Gruyter Mouton), 113–146. doi:10.1515/9783110207569.113.

Bruce, G., Elert, C.-C., Engstrand, O., and Wretling, P. (1999). Phonetics and phonology of the Swedish dialects - a project presentation and a database demonstrator. *Proceedings of the 14th International Congress of Phonetic Sciences, University of California.*, 321–324.

Carpenter, G. A., and Govindarajan, K. K. (1993). "Neural Network and Nearest Neighbor Comparison of Speaker Normalization Methods for Vowel Recognition," in *ICANN '93*, eds. S. Gielen and B. Kappen (London: Springer London), 412–415. doi:10.1007/978-1-4471-2063-6_98.

Chesworth, J., Coté, K., Shaw, C., Williams, S., and Hodge, W. (2003). Effect of phonetic context on women's vowel area. *Canadian Acoustics/Acoustique Canadienne* 31, 20–21.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108, 804–809. doi:10.1016/j.cognition.2008.04.004.

Clopper, C. G. (2009). Computational Methods for Normalizing Acoustic Vowel Data for Talker Differences: Computational Methods for Normalizing Acoustic Vowel Data. *Language and Linguistics Compass* 3, 1430–1442. doi:10.1111/j.1749-818X.2009.00165.x.

Cole, J., Linebaugh, G., Munson, C., and McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics* 38, 167–184. doi:10.1016/j.wocn.2009.08.004.

Crinnion, A. M., Malmskog, B., and Toscano, J. C. (2020). A graph-theoretic approach to identifying acoustic cues for speech sound categorization. *Psychonomic Bulletin & Review* 27, 1104–1125. doi:10.3758/s13423-020-01748-1.

Disner, S. F. (1980). Evaluation of vowel normalization procedures. *The Journal of the Acoustical Society of America* 67, 253–261. doi:10.1121/1.383734.

Eklund, I., and Traunmüller, H. (1997). Comparative Study of Male and Female Whispered and Phonated Versions of the Long Vowels of Swedish. *Phonetica* 54, 1–21. doi:10.1159/000262207.

Elert, C.-C. (1994). "Indelning och gränser inom området för den talade svenskan: En aktuell dialektografi," in *Kulturgränser - myt eller verklighet?* Diabas. (Institutionen för nordiska språk vid Umeå Universitet), 215–228.

Elert, C.-C. (1981). *Ljud och ord i svenskan*. Umeå: Universitetet i Umeå, Almqvist & Wiksell international.

Engstrand, O., Bruce, G., Elert, C.-C., Eriksson, A., and Strangert, E. (2001). Databearbetning i SweDia 2000: Segmentering, transkription och taggning. Version 2.2.

Escudero, P., and Bion, R. A. H. (2007). Modeling vowel normalization and sound perception as sequential processes. *ICPhS* XVI, 1413–1416.

Fabricius, A., Watt, D., and Johnson, D. E. (2009). A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change* 21, 413–435. doi:10.1017/S0954394509990160.

Fant, G. (1983). Feature analysis of Swedish vowels - a revisit. *STL-QPSR* 24, 001–019.

Fant, G. (1975). Non-uniform vowel normalization. *STL-QPSR* 16, 001–019.

Fant, G. (1971). "Notes on the Swedish Vowel System," in *Form and substance: Phonetic and linguistic papers.*, eds. L. Hammerich, R. Jakobson, E. Zwirner, and E. Fischer-Jørgensen (Odense: Andelsbogtrykkeriet).

Fant, G. (2001). Swedish vowels and a new three-parameter model. *TMH-QPSR* 42, 043–049.

Fant, G., Henningsson, G., and Stålhammar, U. (1969). Formant frequencies of Swedish vowels. *STL-QPSR* 10, 026–031.

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review* 116, 752–782.

Flynn, N., and Foulkes, P. (2011). Comparing vowel formant normalization methods. *Proceedings of ICPhS XVII*, 683–686.

Fox, R. A., Flege, J. E., and Munro, M. J. (1995). The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis. *The Journal of the Acoustical Society of America* 97, 2540–2551. doi:10.1121/1.411974.

Fujimura, O. (1967). On the Second Spectral Peak of Front Vowels: A Perceptual Study of the Role of the Second and Third Formants. *Language and Speech* 10, 181–193. doi:10.1177/002383096701000304.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics* 16, 78–80. doi:10.1109/TAU.1968.1161953.

Hadding-Koch, K., and Abramson, A. S. (1964). Duration Versus Spectrum in Swedish Vowels: Some Perceptual Experiments2. *Studia Linguistica* 18, 94–107. doi:10.1111/j.1467-9582.1964.tb00451.x.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristcs of american english vowels. *Journal of the Acoustical Society of America* 97, 3099–3111.

Hillenbrand, J. M., and Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *The Journal of the Acoustical Society of America* 105, 3509–3523. doi:10.1121/1.424676.

Hindle, D. (1978). "Approaches to Vowel Normalization in the Study of Natural Speech," in *Linguistic variation: Models and methods*, ed. D. Sankoff (New York: Academic Press), 161–171.

Hume, E. V., Jaeger, T. F., and Hall, K. (2016). The Message Shapes Phonology.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34, 485–499.

Johnson, K. (2005). Speaker normalization in speech perception. *The Handbook of Speech Perception*. doi:doi:10.1002/9780470757024.ch15.

Johnson, K. (1997). "Speech perception without speaker normalization," in *Talker variability in speech processing*, eds. K. Johnson and W. Mullennix (San Diego: CA: Academic Press), 146–165.

Johnson, K., and Sjerps, M. J. (2021). "Speaker normalization in speech perception," in *The handbook of speech perception* (John Wiley & Sons, Ltd), 145–176. doi:10.1002/9781119184096.ch6.

Joos, M. (1948). Acoustic Phonetics. *Language* 24, 5–136. doi:10.2307/522229.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience* 34, 43–68. doi:10.1080/23273798.2018.1500698.

Kleinschmidt, D. F., and Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review* 122, 148–203. doi:10.1037/a0038695.

Kleinschmidt, D. F., Weatherholtz, K., and Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science* 10, 818–834. doi:10.1111/tops.12331.

Kohn, M. E., and Farrington, C. (2012). Evaluating acoustic speaker normalization algorithms: Evidence from longitudinal child data. *The Journal of the Acoustical Society of America* 131, 2237–2248. doi:10.1121/1.3682061.

Kraljic, T., and Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language* 56, 1–15.

Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). A unified model of categorical effects in consonant and vowel perception. *Psychological Bulletin and Review*, 1681–1712. doi:10.3758/s13423-016-1049-y.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & psychophysics* 50, 93–107. doi:10.3758/BF03212211.

Kulikov, V. (2022). Voice and Emphasis in Arabic Coronal Stops: Evidence for Phonological Compensation. *Language and Speech* 65, 73–104. doi:10.1177/0023830920986821.

Kuronen, M. (2000). *Vokaluttalets akustik i sverigesvenska, finlandssvenska och finska*. Jyväskylä: University of Jyväskylä.

Labov, W. (2010). *Principles of linguistic change. 2: Social factors*. repr. Chichester: Wiley-Blackwell.

Leinonen, T. (2010). An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects. *Groningen Dissertations in Linguistics* 83.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review* 74, 431–461.

Linell, P. (1979). *Psychological reality in phonology: A theoretical study*. Cambridge: Cambridge University Press.

Linell, P. (1978). "Vowel length and consonant length in Swedish word level phonology," in *Nordic prosody: Papers from a symposium* Travaux de l'Institut de Linguistique de Lund., eds. E. Gårding, G. Bruce, and R. Bannert, 123–136.

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

Luce, P. A., and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19, 1–36. doi:10.1097/00003446-199802000-00001.

Luce, R. D. (1959). *Individual choice behavior*. Oxford: John Wiley.

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R., Monto, N., et al. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science* 44, 1–17. doi:10.1111/cogs.12823.

Malinasky, M., Shafiro, V., Moberly, A. C., and Vasil, K. J. (2020). Perception of vowels and consonants in cochlear implant users. *The Journal of the Acoustical Society of America* 148, 2711–2711. doi:10.1121/1.5147511.

McAllister, R., Lubker, J., and Carlson, J. (1974). An EMG study of some characteristics of the Swedish rounded vowels. *Journal of Phonetics* 2, 267–278. doi:10.1016/S0095-4470(19)31297-5.

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology* 18, 1–86.

McMurray, B., Cole, J., and Munson, C. (2011). *Features as an emergent product of computing perceptual cues relative to expectations.*, eds. G. N. Clements and R. Ridouane John Benjamins Publishing Company doi:10.1075/lfab.6.08mcm.

McMurray, B., and Jongman, A. (2016). What comes after /f/?: Prediction in speech derives from data-explanatory processes. *Psychological Science* 27, 43–52. doi:10.1177/0956797615609578.

McMurray, B., and Jongman, A. (2011). What information is necessary for speech categorization?: Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review* 118, 219–246. doi:10.1037/a0022325.What.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of Acoustical Society of America* 85, 22.

Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels*. Indiana.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America* 85, 2088–2113. doi:10.1121/1.397861.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics* 18, 347–373. doi:10.1016/S0095-4470(19)30379-1.

Nearey, T. M. (1983). Vowel-space normalization procedures and phone-preserving transformations of synthetic vowels. *The Journal of the Acoustical Society of America* 74, S17–S17. doi:10.1121/1.2020835.

Nearey, T. M., and Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America* 80, 1297–1308. doi:10.1121/1.394433.

Nearey, T. M., and Assmann, P. F. (2007). "Probabilistic 'sliding template' models for indirect vowel normalization." in *Experimental approaches to phonology*, eds. M.-J. Solé, P. S. Beddor, and M. Ohala (Oxford: Oxford University Press).

Nearey, T. M., and Hogan, J. (1986). "Phonological contrast in experimental phonetics: Relating distributions of measurements production data to perceptual categorization curves," in *Experimental Phonology*, eds. J. J. Ohala and J. Jaeger (New York: Academic Press), 141–161.

Nordstrand, M., Svanfeldt, G., Granström, B., and House, D. (2004). Measurements of articulatory variation in expressive speech for a set of Swedish vowels. *Speech Communication* 44, 187–196. doi:10.1016/j.specom.2004.09.003.

Nordström, P. E., and Lindblom, B. (1975). A normalization procedure for vowel formant data. *Proceedings of ICPhS VIII*.

Norris, D., and McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological review* 115, 357–95. doi:10.1037/0033-295X.115.2.357.

Pardo, J. S., and Remez, R. E. (2006). "Chapter 7 - The Perception of Speech," in *Handbook of Psycholinguistics (Second Edition)*, eds. M. J. Traxler and M. A. Gernsbacher (London: Academic Press), 201–248. doi:10.1016/B978-012369374-7/50008-0.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51, 195–203. doi:10.3758/s13428-018-01193-y.

Persson, A. (2023). Acoustic-perceptual cues to vowel identity in Stockholm Swedish. *Manuscript, Stockholm University*.

Persson, A., and Jaeger, T. F. (2023). The effect of pre-linguistic normalization in vowel perception. *Manuscript, Stockholm University*.

Peterson, G. E. (1961). Parameters of Vowel Quality. *Journal of Speech and Hearing Research* 4, 10–29. doi:10.1044/jshr.0401.10.

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, 175–184.

Pierrehumbert, J. B. (2001). "Exemplar dynamics: Word frequency, lenition and contrast," in *Frequency and the Emergence of Linguistic Structure*, eds. J. Bybee and P. Hopper (John Benjamins), 137–157.

Polka, L., and Bohn, O.-S. (2003). Asymmetries in vowel perception. *Speech Communication* 41, 221–231.

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Riad, T. (2014). *The phonology of Swedish*. Oxford: Oxford University Press.

Richter, C., Feldman, N. H., Salgado, H., and Jansen, A. (2017). Evaluating low-level speech features against human perceptual data. *Transactions of the Association for Computational Linguistics* 5, 425–440. doi:10.1162/tacl_a_00071.

Robb, M. P., and Chen, Y. (2009). Is /h/ phonetically neutral? *Clinical Linguistics & Phonetics* 23, 842–855. doi:10.3109/02699200903247896.

RStudio Team (2020). *RStudio: Integrated development environment for R*. Boston, MA: RStudio, PBC.

Schaeffler, F. (2005). *Phonological quantity in swedish dialects: Typological aspects, phonetic variation and diachronic change*. Umeå: Umeå University, Dep. of philosophy and linguistics.

Sjerps, M. J., Fox, N. P., Johnson, K., and Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications* 10, 1–9. doi:10.1038/s41467-019-10365-z.

Stilp, C. (2020). Acoustic context effects in speech perception. *WIREs Cognitive Science* 11. doi:10.1002/wcs.1517.

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

Syrdal, A. K. (1985). Aspects of a model of the auditory representation of american english vowels. *Speech Communication* 4, 121–135. doi:10.1016/0167-6393(85)90040-8.

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

Tan, M., and Jaeger, T. F. (2022). Listeners adjust prior expectations. *Manuscript, Stockholm University*.

Tan, M., Xie, X., and Jaeger, T. F. (2021). Using rational models to understand experiments on accent adaptation. *Frontiers in Psychology* 12, 1–19. doi:10.3389/fpsyg.2021.676271.

ten Bosch, L., Boves, L., and Ernestus, M. (2022). DIANA, a Process-Oriented Model of Human Auditory Word Recognition. *Brain Sciences* 12, 681. doi:10.3390/brainsci12050681.

Thomas, E. R., and Kendall, T. (2007). NORM: The vowel normalization and plotting suite.

Toscano, J. C., and McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science* 34, 434–464.

Toscano, J. C., and McMurray, B. (2015). The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments. *Language, Cognition and Neuroscience* 30, 529–543. doi:10.1080/23273798.2014.946427.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *The Journal of the Acoustical Society of America* 69, 1465–1475. doi:10.1121/1.385780.

Verbrugge, R. R., and Shankweiler, D. P. (1977). Prosodic information for vowel identity. *Journal of the Acoustical Society of America* 61.

Watt, D., and Fabricius, A. (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*, ed. D. Nelson, 159–173.

Weatherholtz, K., and Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. *Oxford Research Encyclopedia of Linguistics*. doi:10.1093/acrefore/9780199384

Xie, X., Buxó-Lugo, A., and Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in speech prosody. *Cognition* 211. doi:10.1016/j.cognition.2021.104619.

Xie, X., and Jaeger, T. F. (2020). Comparing non-native and native speech: Are L2 productions more variable? *The Journal of the Acoustical Society of America* 147, 3322–3347. doi:10.1121/10.0001141.

Xie, X., Jaeger, T. F., and Kurumada, C. (2022). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational review. doi:10.17605/OSF.IO/Q7GJP.

Yang, J., and Fox, R. A. (2014). Perception of English Vowels by Bilingual Chinese–English and Corresponding Monolingual Listeners. *Language and Speech* 57, 215–237. doi:10.1177/0023830913502774.

Young, N. J., and McGarrah, M. (2021). Forced alignment for Nordic languages: Rapidly constructing a high-quality prototype. *Nordic Journal of Linguistics*, 1–27. doi:10.1017/S033258652100024X.

Zahorian, S. A., and Jagharghi, A. J. (1991). Speaker normalization of static and dynamic vowel spectral features. *The Journal of the Acoustical Society of America* 90, 67–75. doi:10.1121/1.402350.