Nordic Journal of Linguistics (2024), **47**, 73–110
doi:10.1017/S0332586521000275

ARTICLE

The development of linguistic stimuli for the Swedish Situated Phoneme test

Erik Witte[1,2,3], Jonas Ekeroot[4] and Susanne Köbler[1,3]

[1]Faculty of Medicine and Health, School of Health Sciences, SE 701 82, Örebro University [2]Swedish Institute for Disability Research, Örebro University, Örebro, SE 701 82 [3]Audiological Research Centre, Faculty of Medicine and Health, SE 701 82, Örebro University and [4]WS Audiology, Henri-Dunant-Strasse 100, 91058 Erlangen, Germany
Email for correspondence: erik.witte@oru.se

(Received 6 March 2020; revised 8 July 2021; accepted 9 October 2021; first published online 02 December 2021)

Abstract
The speech perception ability of people with hearing loss can be efficiently measured using phonemic-level scoring. We aimed to develop linguistic stimuli suitable for a closed-set phonemic discrimination test in the Swedish language called the Situated Phoneme (SiP) test. The SiP test stimuli that we developed consisted of real monosyllabic words with minimal phonemic contrast, realised by phonetically similar phones. The lexical and sublexical factors of word frequency, phonological neighbourhood density, phonotactic probability, and orthographic transparency were similar between all contrasting words. Each test word was recorded five times by two different speakers, including one male and one female. The accuracy of the test-word recordings was evaluated by 28 normal-hearing subjects in a listening experiment with a silent background using a closed-set design. With a few exceptions, all test words could be correctly discriminated. We discuss the results in terms of content- and construct-validity implications for the Swedish SiP test.

Keywords: hearing impairment; minimal pairs; phonemic discrimination; phonetic distance; speech perception; speech test; Swedish

1. Introduction[1]

Hearing is an integral part of spoken interpersonal communication. The decreased ability to communicate that often accompanies the onset of biologically constituted hearing loss may thus have severe psychological and social consequences for the people affected (see Friberg et al. 2012, Deal et al. 2017). The most common intervention for hearing loss is a HEARING AID fitting. In Sweden today, approximately 400,000 people use hearing aids (Statistics-Sweden 2019). According to Swedish and international standards (ISO 21388 2020), hearing aid fitting should be evaluated using either validated questionnaires or speech-audiometry tests.

As pointed out by Martin et al. (1998), the most common type of SPEECH-AUDIOMETRY TEST used in hearing clinics is a type of WORD-RECOGNITION test often referred to as a phonemically balanced 50-item (PB50) word list. In the context of speech audiometry, the term PHONEMIC BALANCE means that each test list contains the same composition of PHONEMES as the language in general (Lehiste & Peterson 1959). A typical PB50 word list contains only monosyllabic TEST WORDS (TW), which are often preceded by a short carrier phrase. During testing, the words are administered one at a time through auditory presentation. After each TW presentation, the participant must repeat the perceived TW orally. Initially developed in the US in the mid-20th century, PB50 word lists have become standard procedure in hearing clinics in English-speaking countries (with examples such as the American C.I.D. Auditory Test W-22; Hirsh et al. 1952) and many other countries, including Sweden. In Sweden, PB50 test materials still in use today were developed in the 1950s and 1960s (Lidén & Fant 1954). Due to the long history of using PB50 word lists when measuring the word-recognition ability of people with hearing loss, their properties are very well known.

One of these properties is the size of the CRITICAL DIFFERENCES, by which we can determine whether the differences between scores from consecutive test administrations can be considered statistically significant. Over the past 45 years, authors such as Hagerman (1976), Thornton & Raffin (1978), Carney & Schlauch (2007), and Oleson (2010) have made very similar recommendations for the size of the critical differences between word recognition scores. Since the statistical methods used assume that the sampling distribution to which speech-audiometry scores belong can be approximated by BINOMIAL DISTRIBUTIONS, the size of the critical differences depends both on the number of test trials and the proportion of trials with a correct response. Clearly, when using any type of speech-audiometry test to evaluate the benefit of a hearing rehabilitation intervention, it is crucial to be able to tell true score differences from random fluctuations. Unfortunately, however, the critical differences between consecutive test scores are relatively large. For instance, if the score in the first condition, be it without hearing aids, is 70%, the critical difference ranges from 52% to 86%, meaning that the subsequent test score must be better than 86% (or worse than 52%) for the difference to be considered statistically significant, using a confidence level of 95% (Thornton & Raffin 1978). Such large differences are feasible when the test conditions compared involve testing a subject with and without hearing aids (Grunditz & Magnusson 2013). However, when the issue is the evaluation of the benefit from different hearing aids or different hearing aid settings, the expected score differences are much smaller.

Since the size of the critical differences depends partly on the number of items tested, the PB50 word lists can be used to capture smaller differences if the number of items in each consecutive test is increased by running several PB50 word lists. Using 100 TWs in each consecutive test (which is the highest number of items for which the authors mentioned above have calculated critical differences), a score of 70% in the first test requires the test taker to earn a score above 81% (or below 57%) on the second test for the difference to be considered statistically significant. In essence, this means that the benefit provided by the specific hearing intervention under investigation needs to make an additional 12 words (out of 100) perceivable by the subject for the improvement to be statistically significant. Administering each PB50 word list takes between four and five minutes; this is clinically important. Comparing two conditions using 100 words in each condition would therefore take

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

approximately 20 minutes. Comparing even more conditions would likely be feasible only in an exceptional case.

An alternative way to reduce the size of the critical differences, without extending the number of TWs, is by scoring on a phonemic level. If each phoneme within a word list is considered an independent test trial, the number of trials within each list will be considerably higher, which in turn will result in narrower critical difference limits (Olsen et al. 1997). However, this procedure violates two assumptions that need to be fulfilled for the sampling distribution of speech-audiometry scores to be approximated by binomial distributions. The first of these assumptions is that trials within test sessions need to be independent. Due to the PHONOTACTIC structures in languages (see Sigurd 1965), there are relatively strict requirements as to the way phonemes may be combined. Hence, the different phonemes in a word cannot be assumed to form separate independent trials. Second, trials in binomial distributions are required to share exactly the same underlying SUCCESS PROBABILITY. That said, it is well known that the ease of SPEECH PERCEPTION differs largely between different SPEECH SOUNDS, even for normal-hearing people (Miller & Nicely 1955). For people with hearing loss, the audibility of different speech sounds is also affected by the frequency dependence of their hearing sensitivity, which will be highly specific for different individuals (see Bisgaard, Vlaming & Dahlquist 2010). Thus, the success probability of each phoneme in a PB50 word list will not only be very different, but also difficult to predict.

Both the assumption of independence and the assumption of equal success probability may be compensated for mathematically by approximating the observed test scores to binomial distributions of reduced length. (For the assumption of independence, see Boothroyd & Nittrouer 1988; for the assumption of equal success probability, see Hagerman 1976.) Naturally, however, since both the degree of independence of phonemes and the success probability of each phoneme differ depending on the quality of the subject's speech perception (which in turn is the objective of the testing), statistical methods based upon such length reductions should probably be used with caution.

A further alternative, which we elaborate on in the current study, could be the creation of a speech-audiometric test method that, instead of attempting to reduce the size of critical differences, attempts to enlarge the score differences between consecutive test sessions. The basic rationale is that if it can be approximately predetermined which speech sounds cause difficulties in speech perception in a given sound environment, a test that selectively includes only those speech sounds is likely to be more efficient in capturing the benefit from specific hearing rehabilitation interventions than an equivalent test that includes all speech sounds (Woods et al. 2015). If, in addition, the AUDITORY BACKGROUND used in the test is set up to reflect common situations (such as the urban outdoors, offices, day-care centres, or household sound environments), the speech-audiometry results may also be generalisable to situations outside the test booth (see Wagener, Hansen & Ludvigsen 2008, Smeds, Wolters & Rung 2015). With our test approach, we aim to incorporate these two aspects. The technique is therefore referred to as the SITUATED PHONEME (SiP) TEST METHOD. Our study is the first in a series of studies geared toward implementing the SiP test method in an actual Swedish speech-audiometry test, the SIP TEST. The purpose of the SiP test is primarily to evaluate the benefits of hearing

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

rehabilitation interventions, such as hearing aid fittings, cochlear implants, or computer-based analytic auditory training. (For an overview of auditory training, see Henshaw & Ferguson 2013.)

When developing such a test, there are several complex issues to be dealt with, such as how to predetermine which speech sounds are the most difficult for a given subject in a given auditory background, and what type of auditory backgrounds ought to be used. Before solving these issues (which will be addressed in separate publications), we must consider other ones first, such as matters related to the phonology of the target language, the selected test task, the type of linguistic stimulus used, control over confounding variables, and which specific set of speech sounds should be employed.

1.1 Phonological considerations

The creation of a speech test aimed at gauging the perception of different speech sounds naturally requires a precise definition of the term speech sound. Although the existence of the phoneme as a unit of perception has been highly debated within the psycholinguistic literature (see Kazanina, Bowers & Idsardi 2018), we treated the phoneme as a real, but abstract, unit of perception closely tied to a set of actual articulatory and acoustic realisations, referred to as phones. When several phones are realisations of the same phoneme, they are called allophones. This perspective is fully in line with common phonological theory, such as that described for the Swedish language by Riad (2014). Thus, when measuring auditory speech-perception ability on a speech-sound level, the subject always hears a phone. Depending on the construction of the speech test, however, phonemes may also be perceived.

In Central Swedish (often defined as the dialect spoken in the region around Lake Mälaren), the most common phonological analysis assumes nine vowel phonemes and 18 consonant phonemes (Riad 2014). These phonemes are in turn realised in 23 different vowel phones and 24 consonant phones.[2] Even though in stressed syllables, all phonemes have both phonetically long and short allophones, phonemic length is commonly considered to be carried only by vowels. The reason for this is that phonemic length is conditionally distributed between vowels and consonants, whereby in stressed syllables, a short vowel is always followed by a long consonant, and a long vowel always by a short consonant (or no consonant at all). In unstressed syllables, however, all segments are phonetically short. The primary reason that phonemic length is thought to be carried by vowels is that while the long and short consonant allophones differ merely in duration, long and short vowel allophones also show marked quality differences that are retained in unstressed syllables, even though phonetic length is reduced in such positions (Riad 2014). The reason that there are an odd number of vowel phonemes in Central Swedish is that the short allophones of the vowels /e/ and /ɛ/ are neutralised into a single short allophone, pronounced [ɛ̝]. Notwithstanding, several other dialects retain the phonemic contrast between short /e/ and /ɛ/ – realised as [e] and [ɛ̝], respectively – and instead neutralise the short allophones of /ʉ/ and /ø/, pronounced [ɵ] (Riad 2014). Figure 1 and Table 1 present all vowel and consonant phones in the Swedish language, along with their corresponding underlying phonemes.

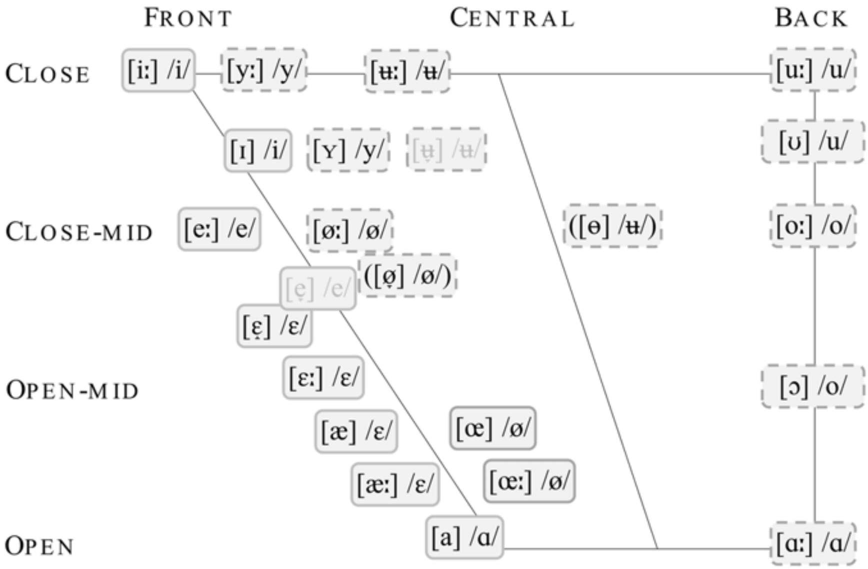

Figure 1. Swedish vowel PHONES and their corresponding underlying PHONEMES. Dashed lines denote rounded phones, while solid lines indicate unrounded phones. All phonemically long vowels also have a length-reduced allophone occurring only in unstressed syllables. The length-reduced allophones are phonetically short but retain the original quality of the long vowels. Phones and phonemes within parentheses undergo neutralisation in some Swedish dialects. The diphthongs [a͡u] and [e͡ʉ], which only occur in a limited number of words, are not shown.

For several phones in Table 1, there are multiple underlying phonemes. With one exception, the RETROFLEX phones are always a result of the COALESCENCE between /r/ (or another retroflex) and one of the dental consonants. For example, /r + t/ are realised as [ʈ]. The exception is [ʂ], which can also be a realisation of an underlying /ʂ/, which in turn has the two conditionally distributed allophones [ʂ] and [ɧ]. In Table 1, phones and phonemes that do not occur in monosyllabic words uttered in isolation are presented in grey.

1.2 Test-task considerations

Existing speech-audiometry tests can be broadly defined in terms of CLOSED- or OPEN-SET formats (Gelfand 2009:261–263). An open-set response format in which the subject can give any response without restrictions can be considered a task of recognition or identification. The process involves the association of an acoustic signal with a set of phonemes that (depending on the type of stimuli used) may correspond to a word stored in the subject's MENTAL LEXICON. (For an overview of theories of speech perception, see Samuel 2010.) Since the number of possible responses in open-set tests is very large, scoring on a phoneme-by-phoneme basis puts rather high demands on the test administrator, who will have to be trained in articulatory phonetics to correctly identify and record the phones uttered by the subject (see Kuk et al. 2010). Even though microphones may be placed near the

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

Table 1. Swedish consonant PHONES and their corresponding primary underlying PHONEMES. Grey colour indicates that the phone, or underlying phoneme, never occurs in monosyllabic words uttered in isolation.

	BILABIAL		LABIODENTAL		DENTAL		RETROFLEX		PALATAL		VELAR		GLOTTAL
	VL[a]	*V*	*VL*	*V*	*VL*	*V*	*VL*	*V*	*VL*	*V*	*VL*	*V*	*VL*
PLOSIVE[b]	[p] /p/	[b] /b/			[t] /t/	[d] /d/	[ʈ] /r+t/	[ɖ] /r+d/			[k] /k/	[g] /g/	
NASAL		[m] /m/		[ɱ] /m+v/		[n] /n/		[ɳ] /r+n/				[ŋ] /ŋ/, /n+g/	
TRILL/FLAP						[r] /r/							
FRICATIVE			[f] /f/	[v] /v/	[s] /s/		[ʂ] /ʂ/, /r+s/		[ɕ] /ɕ/	[j] /j/	[ɧ] /ʂ/		[h] /h/
LATERAL APPROXIMANT						[l] /l/		[ɭ] /r+l/					

[a] VL = VOICELESS, V = VOICED

[b] VOICELESS PLOSIVES can be either ASPIRATED or UNASPIRATED.

subject's mouth, and even if the subject's face is clearly visible to the test administrator, there is a risk that test reliability will be affected due to the test administrator's inevitably subjective interpretation of the response. At present, Swedish audiologists do not have extended training in articulatory phonetics; therefore, the introduction of an open-set phonemic-recognition test for use in everyday evaluations of hearing rehabilitation interventions would likely not be successful.

In contrast, the closed-set format, which allows for selection from a limited number of written response alternatives, is very well suited for everyday clinical use in phonemic-level testing, as it frees the subject from having to deliver an oral response and the test administrator from having to interpret it. The process can be made fully automated using computer software, and requires no further training in articulatory phonetics on the part of the test administrator (see House et al. 1965, Risberg 1976, Feeney & Franks 1982, Greenspan, Bennett & Syrdal 1998, Öster 2006, Nakeva von Mentzer et al. 2018).

While the task in open-set tests is one of word (or phoneme) recognition, the closed-set format morphs the speech-perception task into one of DISCRIMINATION between contrasting response alternatives. In speech tests, such as the SiP test, where the intention is to target testing towards specific phones, this property can potentially be very useful since it allows for the formation of multiple subtests, each consisting of MINIMAL PAIRS, TRIPLETS, or QUADRUPLETS (etc.) that contrast specific sets of TEST PHONES (TP). We refer to such word groups, which manifest minimal phonemic contrast between all member words, as MINIMAL-VARIATION GROUPS (MVGs). When using MVGs as stimuli in a closed-set speech test, testing can be focused on specific phones of interest simply by including only those subtests that contain the phones of interest. If both TWs and response alternatives are presented in random order, the number of trials can, if needed, be increased by multiple repetitions of the selected subtests to gain statistically reliable test scores. For these reasons, the objectives of the SiP test are likely to be best met using a closed-set format.

1.3 Stimulus-type considerations

Both closed- and open-set speech-audiometric tests have been constructed using linguistic forms that exist in the language (i.e. REAL WORDS), as well as those that do not exist (i.e. NON-WORDS) (Gelfand 1998, Rødvik 2008, Paglialonga, Tognola & Grandori 2014, Kollmeier et al. 2015). An advantage of using non-words, be they phonotactically legal or illegal, is that non-words are typically not stored in the subject's mental lexicon and therefore no processes of LEXICAL ACCESS or RETRIEVAL need to be involved. Consequently, the results of such tests may be less prone to variability stemming from factors such as education level or individual variation in the ability to use TOP-DOWN processes, such as PHONEME RESTORATION (see Bashford, Riener & Warren 1992, Samuel 2010). On the other hand, several factors complicate the use of non-words in speech-audiometry tests. For instance, with repeated testing, non-word stimuli will likely undergo a process of MEMORY CONSOLIDATION, by which the subject begins to form mental representations of the non-words. The existence of such processes is indicated by the fact that it is possible to learn to identify non-words present in nonsense babble (Saffran,

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

Newport & Aslin 1996, Mirman et al. 2008), and that repeated exposure to non-words starts to affect other processes such as lexical competition (Gaskell & Dumay 2003).

Non-word perception also seems to be biased toward real words through the Ganong effect (Ganong 1980). If the Ganong effect is not controlled for in a non-word phonemic perception test, it will likely bias the phoneme-perception process in favour of word forms that share similarities with many real words. Finally, an ever-present risk in creating tests based on phonotactically feasible non-words (often referred to as pseudowords) is that such non-words may be changed into real words in the natural course of a language's evolution. Within the Swedish language, for example, some monosyllabic words such as *app* [apː] 'app' or *hen* [hɛ̝nː] 'he/she' were non-words some decades ago, but are very common real words today (Agazzi 2015). These are some reasons why we think that the SiP test should be constructed using real Swedish words rather than non-words. The most important reason is related to the concept of construct validity. In the context of speech audiometry, construct validity means that the object of measurement is actually generalisable to the construct of speech perception (see Shadish, Cook & Campbell 2002:38). In essence, auditory speech perception is a process of deriving meaningful content from an acoustic signal. In order for speech-audiometry tests that are used to evaluate speech perception benefits from hearing rehabilitation interventions to have high levels of construct validity, thus ensuring that the measured construct is truly an aspect of speech perception, the most appropriate choice is to use real words as test stimuli.

1.4 Lexical and sublexical word metrics

Over the years, a number of different lexical and sublexical factors have been seen to influence the speed and accuracy of lexical retrieval. The most prominent of these factors is the word frequency (WF) effect, by which words that are common in the language are easier to perceive (Brysbaert et al. 2015). WF is often defined as occurrences per million words. A psycholinguistically more appropriate metric for the WF effect is the Zipf scale, developed by van Heuven et al. (2014). The Zipf scale is a logarithmic metric ranging from approximately 1 for very uncommon words to 7 for very common words. In contrast to the facilitative nature of the WF effect, the phonological neighbourhood density (PND) effect relates to the process of lexical competition among different – but phonologically similar – words called phonological neighbours. Hence, the presence of many phonological neighbours has an inhibitory effect on lexical retrieval (Luce & Pisoni 1998, Ziegler, Muneaux & Grainger 2003). In a speech audiometry test, such items will be more difficult to recognise than audible words that belong to a sparsely populated phonological neighbourhood (Winkler, Carroll & Holube 2020). The lexical competition offered by different neighbours is not equal, but is instead influenced by, their respective WFs (Luce & Pisoni 1998). A word metric that takes this factor into account is the Zipf-scale weighted phonetic neighbourhood density probability (PNDP), developed for the Swedish language by Witte & Köbler (2019). The PNDP approximates 0 for low-frequency words with many high-frequency neighbours, and rises to 1 for common words with few low-frequency neighbours.

At the sublexical level, PHONOTACTIC PROBABILITY (PP) and ORTHOGRAPHIC TRANSPARENCY (OT) have been seen to facilitate auditory word recognition (Vitevitch & Luce 1998, Dich 2014). PP describes the phonotactic legality of words in a graded manner, rendering higher values for words that contain phonotactic patterns that are common in the given language. While the commonly used metric by Vitevitch & Luce (2004) is supposedly neutral with regard to the target language, Witte & Köbler (2019) invented a similar metric called the NORMALISED STRESS AND SYLLABLE-BASED PP (SSPP), which takes Swedish phonology into account. The word-average SSPP ranges from 0 for words in which all segment transitions are phonotactically illegal, to 1 for words throughout which all segment transitions are highly typical for the language's phonotactic structure.

Finally, OT refers to the probability of phoneme-to-grapheme correspondence or vice versa. Witte & Köbler (2019) calculated several such metrics for the Swedish language, one of which is the GRAPHEME-INITIAL LETTER-TO-PRONUNCIATION OT (GIL2P-OT). The word-average GIL2P-OT describes the level of complexity involved in deriving a pronunciation from a specific string of letters representing a given word. The measure thus aims to quantify the ease of word reading. The word-average GIL2P-OT ranges from 0 for words with very OPAQUE spellings, to 1 for words with very TRANSPARENT spellings.

The Zipf-scale value, PNDP, SSPP, and GIL2P-OT have all been calculated for more than 800,000 Swedish phonetically transcribed words in a freely available psycholinguistic database called the AFC LIST (Witte & Köbler 2019).[3]

Even though the format of the closed-set response test may reduce the biasing influences from various lexical and sublexical properties of specific TWs on the test scores (House et al. 1965, Sommers, Kirk & Pisoni 1997), it is likely that their influence is not eliminated (Foster & Haggard 1987, Clopper, Pisoni & Tierney 2006). Therefore, to minimise the impact of confounding lexical and sublexical properties of the TWs, care should be taken to select contrasting response alternatives, between which the word metrics described above vary as little as possible.

1.5 Test-phone contrasts

In closed-set speech audiometry tests, the number of response alternatives is of some importance. As the number of possible responses is increased, the more difficult the task becomes (Sumby & Pollack 1954). Naturally, the presentation of more response alternatives increases the visual burden of locating the appropriate response, especially if the response alternatives are presented in a new random order in each subsequent trial. On the other hand, with fewer response alternatives, the FLOOR EFFECTS, representing the theoretical average scores from a completely inaudible TW, or simply the chance performance, increase. For instance, using two response alternatives results in a floor effect of 50% (Wichmann & Hill 2001). High levels of chance performance can reduce the overall efficiency of speech-audiometry tests (Yu & Schlauch 2019).

On the other hand, using fewer response alternatives has the practical benefit of increasing the chance of finding appropriate, minimally contrasting real words for

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

which the lexical and sublexical properties described above are relatively similar. If, in addition, the test allows for specific targeting of selected phones, as described above, each set of minimally contrasting response alternatives must be small, lest the phone specificity disappear.

Employing small sets of minimal triplets or quadruplets as response alternatives naturally raises the question of which phones to contrast within each group. Here, several factors suggest that phonetically similar, contrasting phones be used. First, closed-set speech-in-noise tests are considerably easier than open-set tests (Sommers et al. 1997, Kollmeier et al. 2015). To avoid CEILING EFFECTS in the former, the SIGNAL-TO-NOISE RATIO (SNR) may have to be set to ranges with questionable ECOLOGICAL VALIDITY (see Smeds et al. 2015). Contrasting only phonetically similar phones would naturally make the test more difficult, whereby potential ceiling effects would be reduced. Second, a speech-audiometry test – such as the SiP test, which is designed to capture minor changes in the benefits provided by various interventions in hearing rehabilitation – naturally needs to contrast speech segments that are most likely to be confused. Since the likelihood of confusing different phonetic segments depends both on the hearing of each specific subject and on the properties of a potentially present BACKGROUND NOISE (Phatak & Allen 2007, Woods et al. 2015), it is conceivable that phonetically similar phones would be more likely be confused than less similar phones.

However, PHONETIC SIMILARITY and its inverse, PHONETIC DISTANCE (PD), can be determined both acoustically and articulatorily. Since human hearing is essentially non-linear, for both normal-hearing and hearing-impaired individuals, a purely acoustic measure of PD would be rather biased. Several authors have therefore created computational models that can both assimilate the performance of human hearing and calculate measures of PD (Sakoe & Chiba 1978, Holube & Kollmeier 1996, Jürgens & Brand 2009, Mielke 2012). Alternatively, ARTICULATORY FEATURES could be used to determine PD. One difficulty related to calculating PD based on articulation is the fact that articulatory features are essentially categorical along multiple dimensions, for which we know of no proper weighting for the Swedish language. (However, see Kondrak 2003 for a method that could be adapted to Swedish.) Hence, PD may be most appropriately calculated using a computational model based on acoustic waveforms.

1.6 Plans for the SiP test

To summarise, the SiP test method proposed in the current study should present TWs in natural-background sound environments using a closed-set format based on real Swedish words. Each closed set of response alternatives should consist of a small number of real words that contrast phonetically similar phones through minimal phonemic variation, and between which the variations of confounding lexical and sublexical properties are minimised.

To reduce learning effects, both the order in which the TWs are to be presented within each test session and the order in which the written response alternatives are to be presented in each trial should be randomised. In addition, to prevent the subject from learning which response alternatives are most likely the correct responses, all response alternatives should occur as TWs an equal number of times. Therefore,

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

the closed sets of response alternatives also form groups of contrasting TWs, henceforth referred to as TW GROUPS (TWGs). The final SiP test is intended to be presented by means of computer software that mixes the speech signal with background noise, presents the response alternatives on a touch screen, and summarises the test results.

In the current study, we first aimed to develop a set of minimally contrasting, phonetically similar, and psycholinguistically well-controlled TWGs for the SiP test. Our second aim was to evaluate the accuracy of sound recordings of the selected SiP test TWs.

2. Methods

While Section 2.1 concerns the development of TWGs for the SiP test, Section 2.2 entails the development and evaluation of sound recordings of all selected SiP test TWs. Specifically, Section 2.1.1 outlines the technique used to identify sets of candidate TPs, and Section 2.1.2 describes the selection of appropriate sets of minimally contrasting TWs that embody those candidate TPs contrastively. Section 2.2.1 describes the creation of sound recordings of the selected TWs by Swedish speakers, and in Section 2.2.2, a listening experiment is detailed, based on which we evaluated the accuracy of the TW recordings. Because the outcome of this evaluation should not depend on the hearing ability of the participants, but only upon the TW recordings themselves, only normal-hearing subjects participated. The flow chart in Figure 2 outlines the methods used in the current study.

2.1 Development of test-word groups

2.1.1 Selecting candidate test phones

To select contrasting TPs to include in the SiP test, we first calculated a measure of PD between all Swedish phones that occur as contrasting phonemes in Swedish monosyllabic words. The calculation was based on an acoustic analysis of sound recordings of 106 MVGs, which we derived from the AFC list (Witte & Köbler 2019), and which consisted of monosyllabic words of any length. An example of such an MVG would be *far* [fɑːr] 'father', *har* [hɑːr] 'have', and *kar* [kɑːr] 'tub'. The development of these MVGs and the algorithm used to calculate PD are detailed in the Appendix. The 106 MVGs are listed in Supplement 1, the sound recordings of all minimally contrasting words are supplied in Supplement 2, and the resulting PDs between all contrasting minimal pairs within each of the 106 MVGs are available in Supplement 3.

Having calculated PD for all contrasting phones within the 106 MVGs, we computed the average PD for each combination of Swedish phones occurring contrastively in monosyllabic words by averaging across all occurrences of each specific phone combination, as well as across different intrasyllabic positions and phonetic contexts. These average PD values are available in Supplement 4.

Finally, to select sets of suitable TPs to contrast in the SiP test, we identified and clustered around each of the phones included in the PD data derived above the two other phones that were closest in terms of PD. Prior to this clustering, we excluded from the PD data all ZERO-PHONEMES ([∅], indicating the absence of a phoneme in

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

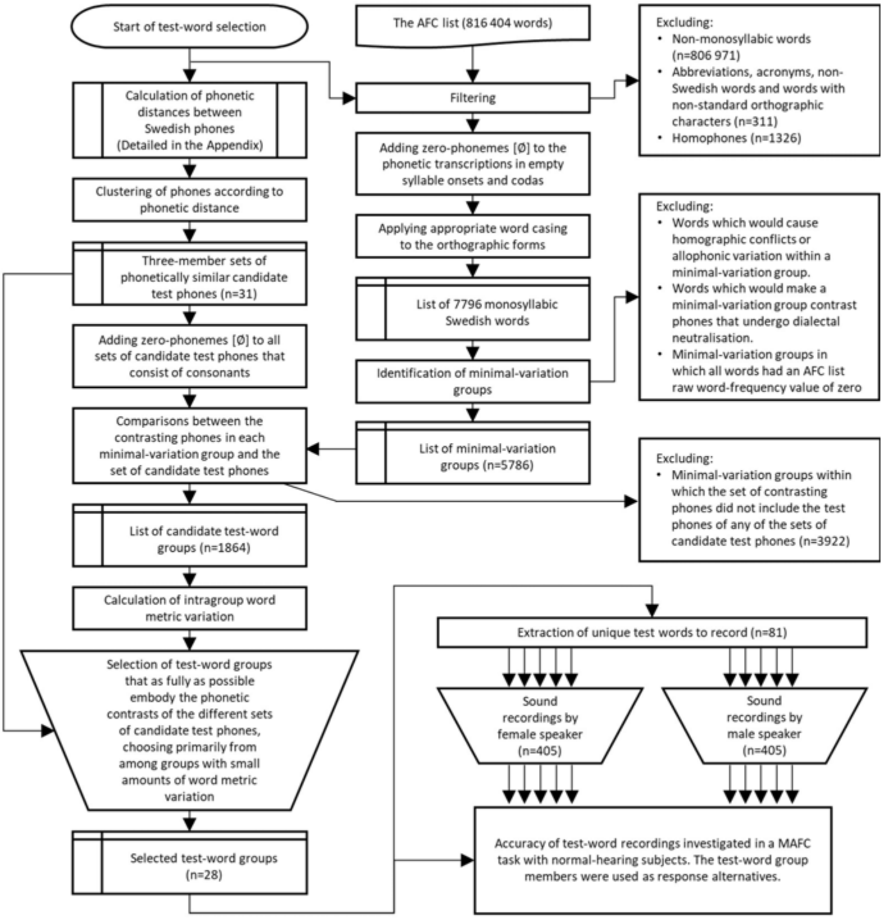
Start of test-word selection
The AFC list (816 404 words)
Excluding:
• Non-monosyllabic words (n=806 971)
• Abbreviations, acronyms, non-Swedish words and words with non-standard orthographic characters (n=311)
• Homophones (n=1326)
Calculation of phonetic distances between Swedish phones (Detailed in the Appendix)
Filtering
Adding zero-phonemes [Ø] to the phonetic transcriptions in empty syllable onsets and codas
Clustering of phones according to phonetic distance
Applying appropriate word casing to the orthographic forms
Excluding:
• Words which would cause homographic conflicts or allophonic variation within a minimal-variation group.
• Words which would make a minimal-variation group contrast phones that undergo dialectal neutralisation.
• Minimal-variation groups in which all words had an AFC list raw word-frequency value of zero
Three-member sets of phonetically similar candidate test phones (n=31)
List of 7796 monosyllabic Swedish words
Adding zero-phonemes [Ø] to all sets of candidate test phones that consist of consonants
Identification of minimal-variation groups
Comparisons between the contrasting phones in each minimal-variation group and the set of candidate test phones
List of minimal-variation groups (n=5786)
Excluding:
• Minimal-variation groups within which the set of contrasting phones did not include the test phones of any of the sets of candidate test phones (n=3922)
List of candidate test-word groups (n=1864)
Calculation of intragroup word metric variation
Extraction of unique test words to record (n=81)
Selection of test-word groups that as fully as possible embody the phonetic contrasts of the different sets of candidate test phones, choosing primarily from among groups with small amounts of word metric variation
Sound recordings by female speaker (n=405)
Sound recordings by male speaker (n=405)
Accuracy of test-word recordings investigated in a MAFC task with normal-hearing subjects. The test-word group members were used as response alternatives.
Selected test-word groups (n=28)

Figure 2. Flow chart describing the steps in the development of linguistic materials for the SiP test in the current study. The AFC list refers to the word-metric database of Witte & Köbler (2019).

empty syllable onsets or codas), diphthongs ([ɑ͡uː] and [e͡ʉː]), and some unusual pronunciations ([ɖː], [ɧː], [vː], [ɳː], [ɭ] and [ŋ]). We added a phone to a cluster only after checking that no allophones – or another phone to which it undergoes NEUTRALISATION in some of the major Swedish dialects (e.g. [ɛ̝]–[e̞], [ɵ]–[ø̞]) – were already added to the same cluster. In the clustering process, we used phonetic transcriptions without length markings[4] to cluster long and short allophones together based on their average PDs. After clustering, we again retrieved both long and short allophones by duplicating each cluster into long and short variants; we did so by reapplying phonetic length characters where appropriate. As a consequence, some of the resulting clusters (available in Supplement 5) could be expressed as containing either long or short phones, lending greater flexibility in identifying sets of minimally contrasting Swedish words within which the various sets of candidate TPs should occur.

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

2.1.2 Test-word selection

Having derived appropriate sets of candidate TPs to include in the SiP test, the next step was to identify appropriate real-word MVGs that contrast those phones phonemically.

To identify such MVGs, we used the monosyllabic words from the AFC list (Witte & Köbler 2019) described above. Before searching the AFC list for appropriate MVGs, we excluded all non-monosyllabic words, words marked as abbreviations, acronyms, foreign (i.e. non-Swedish) words, or words with non-standard orthographic characters. We defined non-standard orthographic characters as not belonging to the following set of standard Swedish orthographic characters: *a*, *b*, *c*, *d*, *e*, *é*, *f*, *g*, *h*, *i*, *j*, *k*, *l*, *m*, *n*, *o*, *p*, *q*, *r*, *s*, *t*, *u*, *v*, *w*, *x*, *y*, *z*, *å*, *ä*, *ö*. We also made sure that no two words had the same phonetic transcription by selectively removing homophones, which we defined as words with the same pronunciation but different spellings.[5]

Then, we added zero-phonemes [∅] to the phonetic transcriptions of all empty syllable onsets or codas in the remaining words. The reason for marking zero-phonemes in this way was to make it possible to test someone's ability to discriminate between phones that are difficult to hear due to hearing loss and the true absence of a phone. Thus, we henceforth treated zero-phonemes as separate, independent phonemes. Since the TWs in the selected MVGs will be presented as written response alternatives, they should appear in the most common CASE. All orthographic forms in the AFC list are represented in the LOWER CASE. However, the AFC list also contains an UPPERCASE field holding the normal proportion of word-initial UPPER CASE usage for each word.[6] Therefore, to ensure appropriate casing of the SiP test TWs, we changed the initial letter of the remaining words to upper case if the AFC-list UpperCase field equalled 100%, or the most common AFC-list word-class assignment was a proper noun and the AFC-list UpperCase field exceeded 70%.

To identify MVGs among the remaining words, we systematically compared each word with every other word, clustering words for which the phonetic transcription was identical, with the exception of exactly one phone. When assembling MVGs, we ensured that no allophones, dialectally neutralised phones, or homographs were included in the same group.[7] We defined homographs as words with identical spelling but different pronunciations.

We then compared each identified MVG to all sets of candidate TPs identified in Section 2.1, and excluded groups that did not contain as phone contrasts any of the sets of candidate TPs from Section 2.1. For an MVG to be considered for further analysis, it had to contrast all phones in at least one of the sets of candidate TPs. If the MVG also contained other phone contrasts (i.e. the group had more than three or four members), we did not entirely remove those excess words, but instead kept them alongside the MVG. We refer to such words as GHOST WORDS, since they occur in the language but will not be available as response alternatives in the final set of TWGs. The reason to keep those words in the data was to enable analysis of their influence on the final SiP test scores, should there be any. Having identified all MVGs, we removed groups in which all words had a raw WF of zero in the AFC list (i.e. not occurring in the corpora upon which the AFC-list WF data were based).

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

To select the most suitable MVGs to include in the SiP test, we sought to minimise the INTRAGROUP VARIATION in WF, PND, PP, and OT. For each candidate group, we thus calculated the COEFFICIENTS OF VARIATION (CV, i.e. the STANDARD DEVIATION, SD, divided by the MEAN) for the Zipf-scale value, the PNDP, the word-average SSPP, and the word-average GIL2P-OT metrics, as derived from the AFC list, and then rank-ordered those CV values. To arrive at a single value of intragroup word-metric variation for each candidate TWG, we computed its average rank across all four word metrics. For each set of candidate TPs identified in Section 2.1, we then attempted to select at least one TWG among those with relatively low intragroup word metric variation. In cases when we could not find any appropriate MVG containing the desired set of candidate TPs, one phone was allowed to be dropped from the set as long as the number of member words did not fall below three. If we still could not identify any appropriate MVG, we removed that particular set of candidate TPs from the material. To ensure proper coverage of high-frequency speech sounds, two sets of candidate TPs containing VOICELESS FRICATIVES were each represented by two different MVGs.

In addition to the actual TWGs, we selected one MVG intended for use in a practice test. This group consisted of contrasting vowel sounds.

2.2 Development of test-word recordings

2.2.1 Sound recordings

Having chosen a set of candidate words for the SiP test, the next step was to create and validate the accuracy of sound recordings of these words. For each TW, five recordings were made by two native speakers of Central Swedish, one male (the second author) and one female. During these recordings, speech-weighted noise was presented to the talkers via headphones at approximately 65 dB C. The purpose of this noise was to trigger a LOMBARD EFFECT (see Van Summers et al. 1988) in the talkers, naturally giving them a relatively raised vocal effort, as the TWs in the final SiP test will be presented in background noise. Prior to each TW recording, a sound file containing the TW recorded by the first author was presented to the talker via headphones. We refer to these as PROTOTYPE RECORDINGS. The purpose of presenting these prototype recordings was to attain a similar PITCH, INTONATION, and INTENSITY in all TW recordings. For the same purpose, all prototype words were in turn recorded using one single monosyllabic word as a prototype. Another purpose of using the prototype recordings was to free the talkers from concentrating on reading the phonetic transcriptions of the TWs, whereby the risk of erroneous recordings could be minimised. Instead, the SPELLING of the TW was presented to the talker on a computer screen. For the talkers to hear the prototype recordings clearly, the speech-weighted noise was temporarily turned off while the prototype recordings were played. Each word could be re-recorded several times if the speaker or the first author (who carefully monitored the entire recording process from a control room) was not satisfied with the outcome. The speakers themselves decided when to move on to the next word by pressing a button, and there was no time limit between consecutive words.

The recordings took place in an anechoic chamber at the AUDIOLOGICAL RESEARCH CENTRE in Örebro. A Neumann TLM 107 condenser microphone[8] with

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

a pop screen was used, together with an external sound card (RME Fireface UC) connected to a host PC, running custom-written recording software.[9] The speech-weighted noise, along with the prototype recordings, was presented to the talkers through supra-aural headphones.[10]

2.2.2 Investigation of test-word recording accuracy

To investigate the accuracy of the TW recordings, 28 normal-hearing subjects (16 females, 12 males; 24–72 years old; M = 40.3, SD = 13.9) participated in a listening experiment. All subjects were native speakers of Swedish and had normal hearing, as determined by pure-tone audiometric screening (for the audiometric frequencies 125–8000 Hz, using as screening level the 75th percentile of the age and sex cohort of each subject as defined in ISO 7029 2000), as well as by normal scores (< 17) on the Swedish Hearing Handicap Inventory for the Elderly (HHIE, Öberg, Lunner & Andersson 2007). None of the participants had diagnosed dyslexia or severe visual impairment.

The listening experiment used a closed-set design, similar to that planned for the final version of the SiP test described in the introduction, but not entirely identical. In the current listening experiment, all participants took part in a single test session in which they were seated in front of a touch screen and asked to listen for short Swedish words. In the test sessions, SiP test TWs were presented from a single loudspeaker (Genelec 8040B) in front of the participants at an average sound level of 62.35 dB SPL, representing a normal vocal effort level (ANSI-S3.5 1997).[11] No background noise was used. After each auditory TW presentation, the participant was asked to indicate which word he or she had perceived by selecting from a closed set of alternatives appearing in written form on the touch screen 0.5 seconds after the end of the auditory TW presentation. The alternatives were always all members of the MVG to which the auditorily presented TW belonged. The participants were asked to respond as quickly and accurately as possible and to make a guess if in doubt. If the participants had not answered within four seconds, counting from the presentation of the response alternatives, a missing response was recorded and the test was continued. The interstimulus intervals (not counting the response time) were randomised between consecutive trials within a range of 0.9 to 1.5 seconds. All TWs and response alternatives were presented in a random order. When an incorrect or missing response occurred, two additional presentations of that TW recording were inserted in a random position among the remaining test trials. Each participant was presented with one recording by each speaker of each member word in all TWGs at least once. Thus, at least 182 test trials were presented in each test session. The listening experiment took place in an anechoic chamber, and each test session lasted for approximately 15 minutes.

Since the TW recordings were presented without background noise to normal-hearing subjects, we expected that the participants would make no errors, except those occurring due to lapses of attention or random mistakes. When analysing the results of the listening experiment, we therefore ignored single errors, defined as errors occurring only once per TW, SiP test voice, and test session. Instead, we focused on recurring errors and thus calculated, separately for each TW and SiP test voice, the proportion out of the 28 test sessions in which repeated errors (defined

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

as errors occurring more than once per TW and test session) had occurred. TW recordings for which no such repeated errors occurred were considered to have a high level of accuracy. We did not perform any further statistical analysis.

2.2.3 Ethical considerations

We conducted the listening experiment described above in accordance with the Declaration of Helsinki (World Medical Association 2013).[12]

3. Results

3.1 Test-word groups

Table 2 depicts the results of the selection of TWGs for the SiP test. The sets of contrasting phones manifested in the Swedish MVGs in Table 2 embody most of the 31 sets of candidate TPs identified in Section 2.1.1 (and presented in Supplement 5). However, for four of the sets of candidate TPs (i.e. [d ɖ ɳ], [b d ɖ], [d ɡ ɳ] and [ɔ œ ʊ]), we could not identify any appropriate MVG. For the same reason, three sets of candidate TPs had to be reduced by dropping [m] from the set [m n ɳ], [ɖ] from the set [ɖ n ɳ], and [ɳ] from the set [l n ɳ]. In all cases, three-member TWGs were formed by adding zero-phonemes. In total, we selected 28 MVGs as TWGs for the SiP test. Supplement 6 portrays the ghost words for each selected TWG.

The distributions of contrasting phones in the selected TWGs are presented in Figure 3 for vowels and Figure 4 for (length-invariant) consonants. When consonant length is reduced, and with the exception of the already excluded diphthongs and the retroflex lateral [ɭ], [ɖ] is the only phone allowed in Swedish monosyllabic words,[13] which we did not included as a TP in any of the TWGs selected for the SiP test.

Figures 5–7 present the sets of contrasting phones within the selected TWGs in terms of articulatory features. The left and right panels in Figure 5 indicate groups of long and short vowels, respectively.

Figure 6 outlines consonant groups within which the phones differ only across the dimension of PLACE OF ARTICULATION. In contrast, Figure 7 depicts phones that differ along more or other articulatory dimensions than place of articulation. Figures 6 and 7 do not show the zero-phonemes [∅] included in several of the consonant groups (see Table 2).

Figures 8–11 present statistics on the four types of word metrics for which we minimised variation during the TW selection process, namely, WF expressed by the Zipf-scale value (Figure 8), PND using the PNDP metric (Figure 9), PP using the word-average SSPP metric (Figure 10), and finally, OT using the word-average GIL2P-OT metric (Figure 11).[14] In these figures, word-metric values are shown for all words within each selected TWG, along with the word-metric variation in each TWG. Values for specific words are indicated by black points, identified by the corresponding contrasting phone. Mean values for each TWG are denoted by black crosses. Vowel groups and consonant groups are depicted in separate facets within each figure. The data in each facet are ordered so that the intragroup ranges (given in parentheses) for the different word metric values decrease from the top to bottom of each facet.

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

Table 2. Spellings, phonetic transcriptions, and contrasting phones in the MINIMAL-VARIATION GROUPS selected as TEST-WORD GROUPS for the SiP test. The categories LONG and SHORT correspond to the phonetic length of the contrasting phones.

		SPELLING	PHONETIC TRANSCRIPTION	CONTRASTING PHONES
VOWELS	SHORT	*sitt, sytt, sött*	[sɪtː], [sʏtː], [sø̞tː]	[ɪ], [ʏ], [ø̞]
		sätt, sitt, sytt	[sɛ̝tː], [sɪtː], [sʏtː]	[ɛ̝], [ɪ], [ʏ]
		satt, sätt, sött	[satː], [sɛ̝tː], [sø̞tː]	[a], [ɛ̝], [ø̞]
		mark, märk, mörk	[marːk], [mærːk], [mœrːk]	[a], [æ], [œ]
		bland, blond, blund	[blanːd], [blɔnːd], [blɵnːd]	[a], [ɔ], [ɵ]
		sarg, sorg, sörj	[sarːj̥], [sɔrːj̥], [sœrːj̥]	[a], [ɔ], [œ]
		rått, rott, rött	[rɔtː], [rʊtː], [rø̞tː]	[ɔ], [ʊ], [ø̞]
	LONG	*pir, pur, pyr*	[piːr], [pʉːr], [pyːr]	[iː], [ʉː], [yː]
		red, räd, Ryd	[reːd], [rɛːd], [ryːd]	[eː], [ɛː], [yː]
		Klas, kläs, klös	[klɑːs], [klɛːs], [kløːs]	[ɑː], [ɛː], [øː]
		mas, mås, mös	[mɑːs], [moːs], [møːs]	[ɑː], [oː], [øː]
		mår, mor, mur	[moːr], [muːr], [mʉːr]	[oː], [uː], [ʉː]
CONSONANTS	SHORT	*hy, hyf, hys, hyrs*	[hyː∅], [hyːf], [hyːs], [hyːʂ]	[∅], [f], [s], [ʂ]
		arm, farm, charm, larm	[∅arːm], [farːm], [ɧarːm], [larːm]	[∅], [f], [ɧ], [l]
		yr, fyr, skyr, syr	[∅yːr], [fyːr], [ɧyːr], [syːr]	[∅], [f], [ɧ], [s]
		å, få, sjå, så	[∅oː∅], [foː∅], [ɧoː∅], [soː∅]	[∅], [f], [ɧ], [s]
		all, hall, pall, tall	[∅alː], [halː], [palː], [talː]	[∅], [h], [p], [t]
		il, kil, fil, sil	[∅iːl], [ɕiːl], [fiːl], [siːl]	[∅], [ɕ], [f], [s]
		ur, bur, dur, mur	[∅ʉːr], [bʉːr], [dʉːr], [mʉːr]	[∅], [b], [d], [m]
		ko, kon, korn	[kuː∅], [kuːn], [kuːɳ]	[∅], [n], [ɳ]
		ed, led, ned	[∅eːd], [leːd], [neːd]	[∅], [l], [n]
		kval, kvarn, kvar	[kvɑːl], [kvɑːɳ], [kvɑːr]	[l], [ɳ], [r]
		kval, kvarn, kvav	[kvɑːl], [kvɑːɳ], [kvɑːv]	[l], [ɳ], [v]
	LONG	*tuff, tuss, tusch*	[tɵfː], [tɵsː], [tɵʂː]	[fː], [sː], [ʂː]
		sopp, sått, sort	[sɔpː], [sɔtː], [sɔʈː]	[pː], [tː], [ʈː]
		sock, sått, sort	[sɔkː], [sɔtː], [sɔʈː]	[kː], [tː], [ʈː]
		tugg, tum, tung	[tɵgː], [tɵmː], [tɵŋː]	[gː], [mː], [ŋː]
		paj, pall, pang	[paj̥ː], [palː], [paŋː]	[j̥ː], [lː], [ŋː]

Note: The symbol [∅] (ZERO-PHONEME) corresponds to the absence of a phoneme.

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

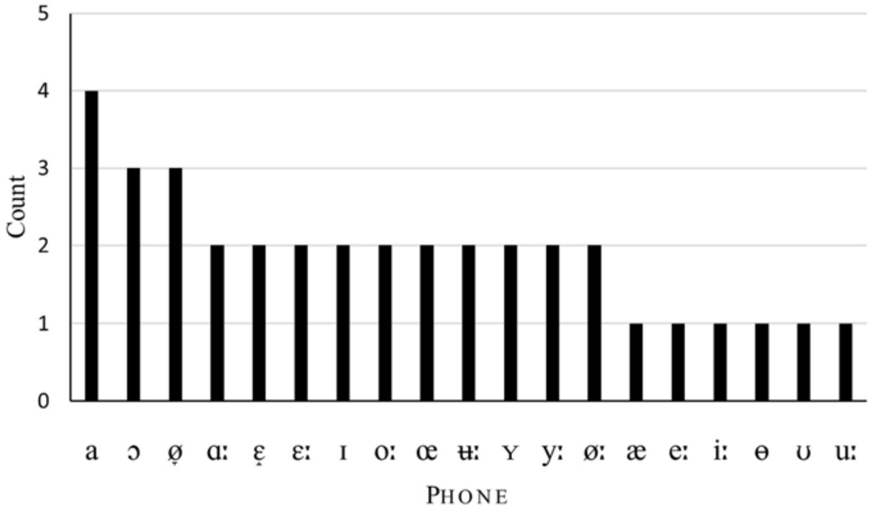

Figure 3. The distribution of Swedish vowels included among the TEST PHONES in the TEST-WORD GROUPS selected for the SiP test.

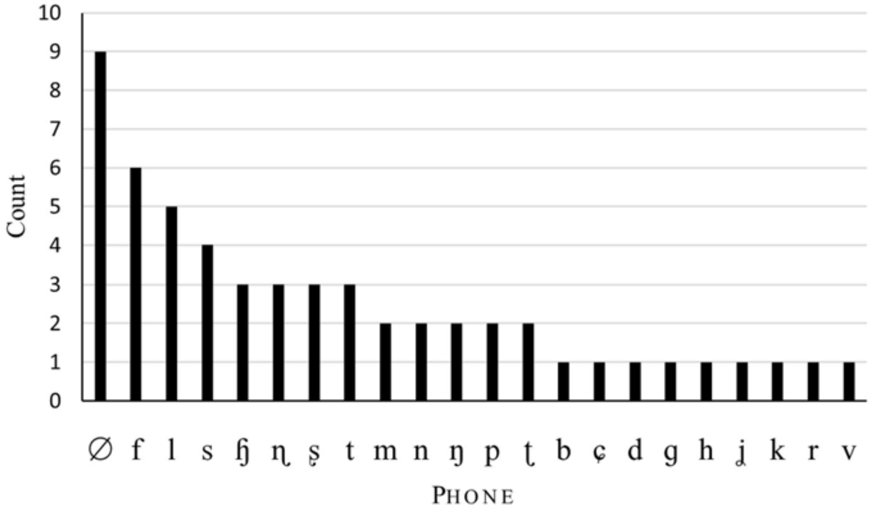

Figure 4. The distribution of Swedish consonants (invariant of length) included among the TEST PHONES in the TEST-WORD GROUPS selected for the SiP test.

As seen in Figure 8, the WF range within most TWGs is below two units on the Zipf scale. Two groups, [hyː∅], [hyːf], [hyːs], [hyːʂ] and [∅oː∅], [foː∅], [ɧoː∅], [soː∅], contain extremely common or uncommon words. Most other groups contain words that range between 2 and 5.5 on the Zipf scale. The variation in PND, as outlined in Figure 9, is relatively low for most TWGs. However, a few groups show a much wider variation, such as [marːk], [mærːk], [mœrːk], and [kvɑːl], [kvɑːɳ],

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

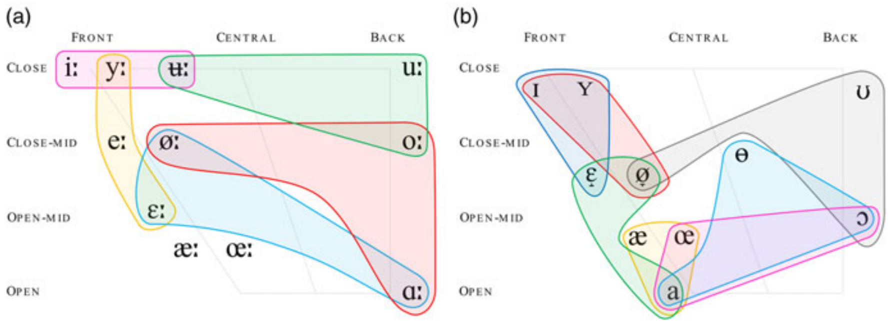

Figure 5. The sets of contrasting vowel phones within the selected TEST-WORD GROUPS are presented in terms of articulatory features. The left (a) and right (b) panels present long and short vowels, respectively.

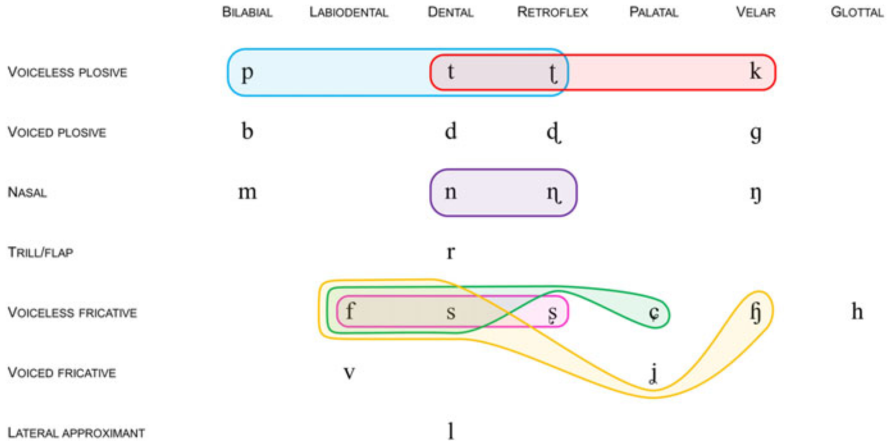

Figure 6. SiP test TEST-PHONE contrasts that differ along the articulatory dimension PLACE OF ARTICULATION. The areas represent different TEST-WORD GROUPS in the SiP test. Zero-phonemes [∅] are not shown.

[kvɑːv]. The pattern is nearly the same in Figure 10, where the variation in PP is generally low in most cases, especially among the vowel groups. Nevertheless, some consonant groups indicate rather large intragroup variation in PP. As with WF, the most extreme values are again denoted by the groups [hyː∅], [hyːf], [hyːs], [hyːʂ] and [∅oː∅], [foː∅], [ɧoː∅], [soː∅]. Finally, in Figure 11, all groups contain words with relatively high OT, with values of GIL2P-OT approaching or exceeding 0.9 in all cases.

In Figure 12, the left panel (a) presents the distributions of the Zipf-scale value, PNDP, word-average SSPP, and word-average GIL2P-OT among all monosyllabic words of the AFC list, along with their corresponding means and SDs. In the right panel (b) of Figure 12, the distribution of SDs within the selected TWGs are shown for the same four word metrics. The mean of each distribution in Figure 12b is also given within each facet. For instance, the mean of the SD of the Zipf-scale value in Figure 12b is 0.82. The SD of the Zipf-scale value among all Swedish monosyllabic

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

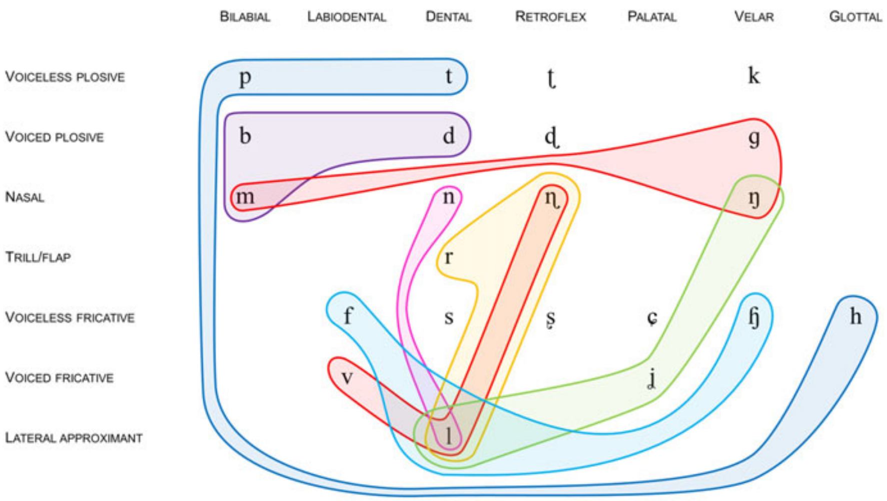

Figure 7. SiP test test-phone contrasts differing along more or other articulatory dimensions than place of articulation. The areas represent different test-word groups in the SiP test. Zero-phonemes [∅] are not shown.

words in Figure 12a is 1.32. Thus, the variation in WF within the selected TWGs is on average only 62% (0.82/1.32) of the typical variation seen among Swedish monosyllabic words. Similar relations are also observed with the other three word metrics studied, for which the average variation in the selected TWG is 14%, 50%, and 29% of the variation among all Swedish monosyllabic words for PND (PNDP), PP (word-average SSPP), and OT (word-average GIL2P-OT), respectively. Thus, the variation in all four word metric types is generally lower within the selected TWGs than among Swedish monosyllabic words in general.

The four distributions depicted in Figure 12b are not mutually comparable since the original scale differs between the metrics. To transform the intragroup variation in the different word metrics to the same (unit-less) scale, we computed the CV for each TWG and word metric. The distribution of CV for the selected TWGs is portrayed separately for the four word metrics studied in Figure 13. Clearly, the mean degree of remaining intragroup variation is much larger for the WF (Zipf-scale) and PND (PNDP) metrics than for the metrics of PP (word-average SSPP) and OT (word-average GIL2P-OT).

3.2 Accuracy of test-word recordings

Contrary to our expectations, the results of the listening experiment, in which we investigated the accuracy of the TW recordings, indicated that repeated incorrect responses occurred for six separate TWs. For these words, Table 3 displays the proportion of test sessions with repeated errors, along with the most common erroneous response for each word. With the male voice recordings, the word *pyr* [pyːr] 'smoulder' was repeatedly confused with the word *pir* [piːr] 'pier' in five of the 28 test sessions (18%). For the female voice, the same confusion occurred repeatedly

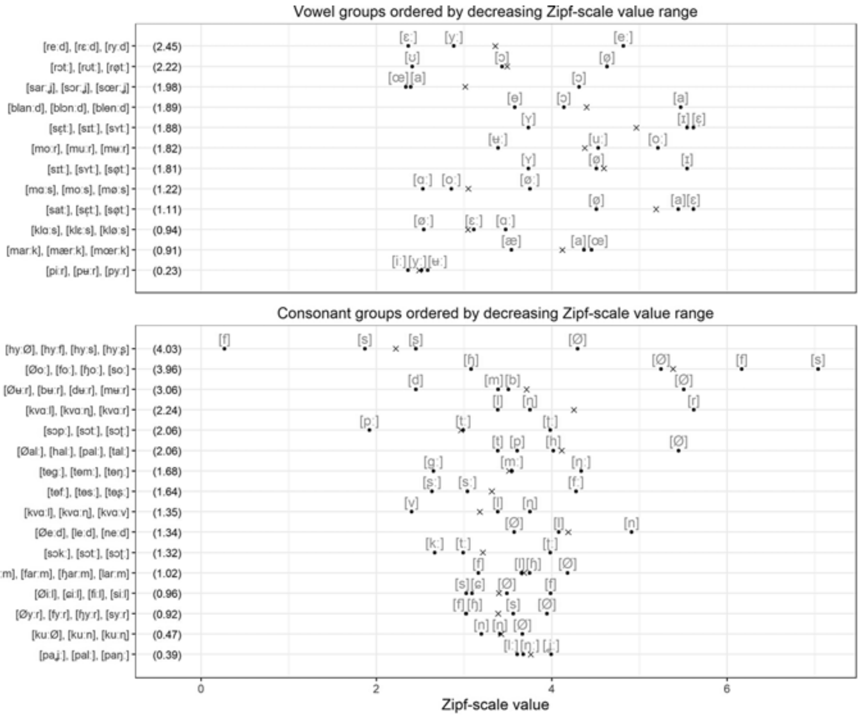

Figure 8. The WORD FREQUENCY (WF), as given by the Zipf-scale metric, for each TEST WORD included in the SiP test material. Mean values for each TEST-WORD GROUP (TWG) are indicated by black crosses, and values for specific words are denoted by black points labelled with the corresponding contrasting phone. TWGs are presented separately for vowel and consonant groups in descending order of intragroup WF range, indicated by the numbers in parentheses.

in 20 out of the 28 test sessions (71%). The opposite confusion of *pir* [piːr] for *pyr* [pyːr] never occurred. In general, the female voice generated more errors than the male voice.

Of the TWs not included in Table 3, most words were correctly responded to at their first presentation. The only exceptions were *charm* [ɧarːm] 'charm' and *rått* [rɔtː] 'raw', which had single errors in one test session each. Supplement 7 contains a summary of the scores for each TW and TW recording separately.

We have made all sound recordings of the SiP test TWs available online at https://osf.io/y4nqb under a Creative Commons Attribution Non-Commercial 4.0 International (CC BY-NC 4.0) licence (https://creativecommons.org/licenses/by-nc/4.0/).

4. Discussion

4.1 Content validity

The CONTENT VALIDITY of a test refers to the extent to which its items adequately cover the construct under investigation (Streiner & Norman 2008:24). For the Swedish SiP test to have a high level of content validity, it will need to include

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

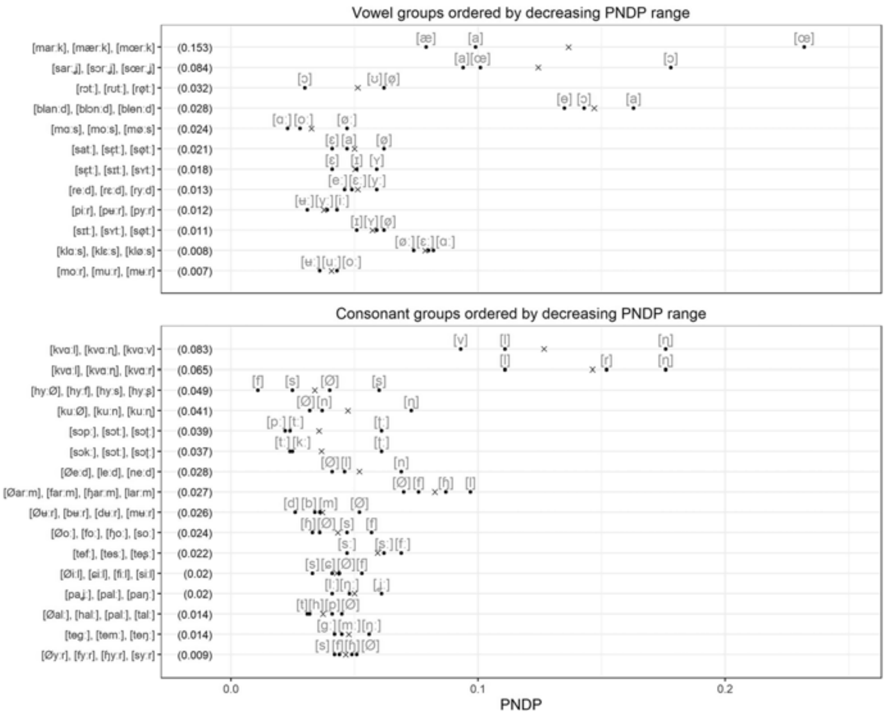

Figure 9. The PHONOLOGICAL NEIGHBOURHOOD DENSITY (PND), as given by the ZIPF-SCALE WEIGHTED PHONETIC NEIGHBOURHOOD DENSITY PROBABILITY (PNDP) metric, for each TEST WORD included in the SiP test material. Mean values for each TEST-WORD GROUP (TWG) are indicated by black crosses, and values for specific words are denoted by black points labelled with the corresponding contrasting phone. TWGs are presented separately for vowel and consonant groups in descending order of intragroup PND range, indicated by the numbers in parentheses.

the full range of Swedish phones as TPs, each of which needs to be contrasted to other phonetically similar phones.

4.1.1 Selected test phones

With a few exceptions, all speech sounds that can be manifested in Swedish monosyllabic words, as specified by the transcription convention of the AFC list (Witte & Köbler 2019), were represented among the TPs in one or more of the TWGs selected for the SiP test. However, all Swedish monosyllabic words uttered in isolation are stressed and thus always contain at least one phonemically long segment. In effect, this means that we have not included any realisations of the underlyingly long segments which undergo length reduction in unstressed syllables in the SiP test material. To include such segments, non-monosyllabic words would also need to be added to the material. In addition, the Swedish language has both phonemic length and phonemic PITCH ACCENT, neither of which is contrasted in the SiP test material. The reason we did not contrast pitch accent also follows from the fact that the SiP

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

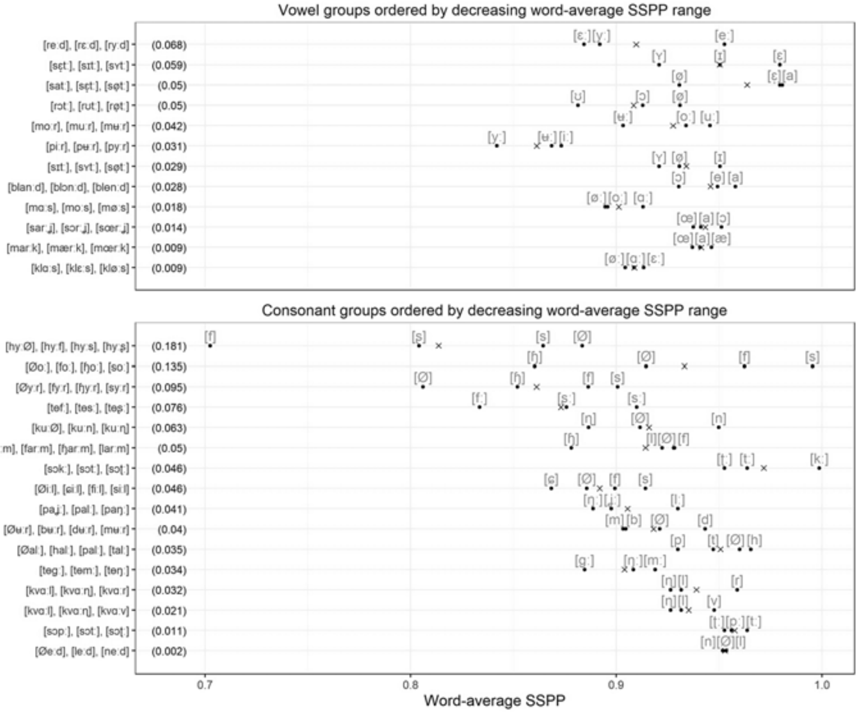

Figure 10. The PHONOTACTIC PROBABILITY (PP), as given by the word-average NORMALISED STRESS AND SYLLABLE-BASED PP (SSPP) value, for each TEST WORD included in the SiP test material. Mean values for each TEST-WORD GROUP (TWG) are signalled by black crosses, and values for specific words are indicated by black points labelled with the corresponding contrasting phone. TWGs are presented separately for vowel and consonant groups in descending order of intragroup PP range, indicated by the numbers in parentheses.

test only contains monosyllabic words; because contrasting pitch accents in Swedish always span across two syllables, all Swedish monosyllabic words have the same pitch accent. We did not contrast phonemic length because phonetically, length in the Swedish language is complementarily distributed between juxtaposed consonants and vowels (Riad 2014), and therefore does not express true minimal phonemic variation.

In addition to these restrictions, further limitations of the material result from the fact that we have not included manifestations of the selected phonemic contrasts in every possible intrasyllabic location and different phonetic context. The well-known COARTICULATORY effects by which both perception and the acoustic content of speech sounds are influenced by the properties of the surrounding phonetic segments (see Diehl, Lotto & Holt 2004) are hence not accounted for in the SiP test. However, attempting to control for all occurring coarticulatory effects would require the SiP test to include many more MVGs, which would ultimately lead to very lengthy testing times and possible listening fatigue (see Pedersen & Zacharov 2018).

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

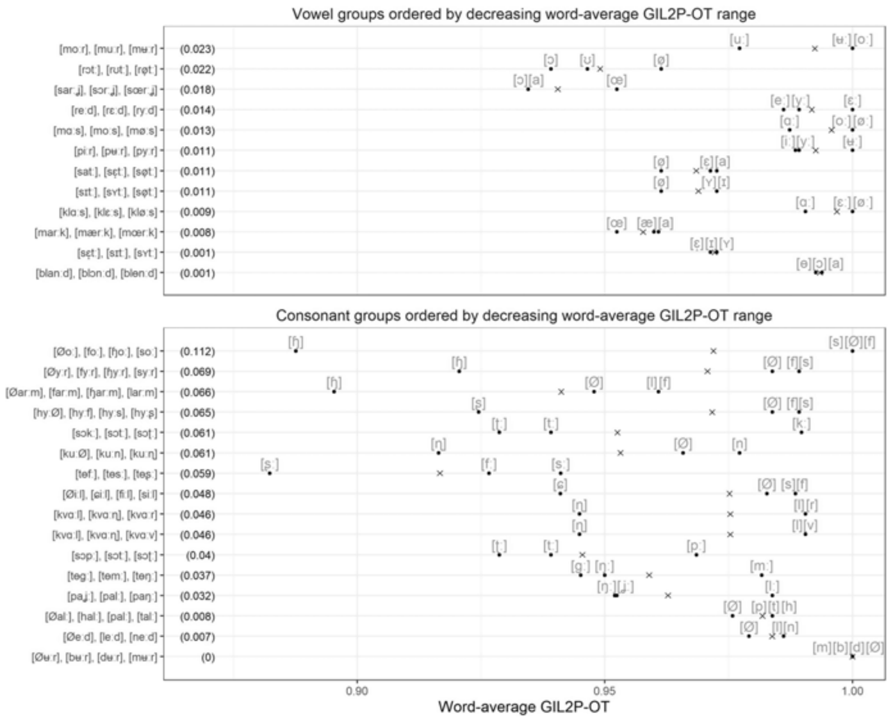

Figure 11. The ORTHOGRAPHIC TRANSPARENCY (OT), as given by the word average GRAPHEME-INITIAL LETTER-TO-PRONUNCIATION ORTHOGRAPHIC TRANSPARENCY (GIL2P-OT) metric, for each TEST WORD included in the SiP test material. Mean values for each TEST-WORD GROUP (TWG) are denoted by black crosses, and values for specific words are represented by black points labelled with the corresponding contrasting phone. TWGs are presented separately for vowel and consonant groups in descending order of intragroup OT range, indicated by the numbers in parentheses.

4.1.2 Selected phone contrasts

The PD metric, which laid the foundation for our selection of TPs, was based on computational analyses of acoustic realisations of Swedish phonemes. An alternative way of selecting contrasting phones for the SiP test could have been grounded in experimentally obtained confusion patterns derived from people with various degrees of hearing loss in diverse types of auditory backgrounds (Miller & Nicely 1955, Välimaa et al. 2002, Phatak & Allen 2007, Rødvik 2008). Although the metric of PD employed here was rather computationally complex, it has the advantage of not being based on behavioural responses. The computational model entails a purely stimulus-driven, BOTTOM-UP process. As such, it does not suffer from the problems of bias stemming from the top-down lexical and sublexical influences discussed in the introduction. Nor does the computational model employed here depend on proper weighting of articulatory dimensions (compare Kondrak 2003) since no such features are needed. The latter point raises the question as to what degree the same sets of candidate phones could have been determined directly from the vowel quadrant and consonant charts presented in Figure 1

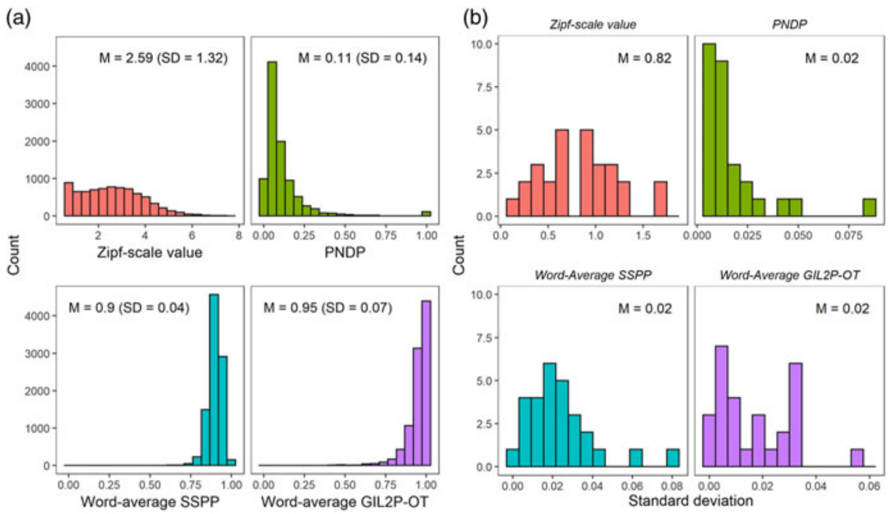

Figure 12. The left panel (a) shows the distribution of WORD FREQUENCY (Zipf-scale value), PHONOLOGICAL NEIGHBOURHOOD DENSITY (PNDP), PHONOTACTIC PROBABILITY (word-average SSPP), and ORTHOGRAPHIC TRANSPARENCY (word-average GIL2P-OT) among all monosyllabic words in the AFC list. Corresponding means (M) and standard deviations (SD) are given in each facet. The right panel (b) displays the distribution of SDs for the same word metrics within the TEST-WORD GROUPS (TWG) selected for the SiP test. The mean of each distribution in panel b is given in the corresponding facet. As the mean SDs in panel (b) are all below the corresponding SDs in panel a, the variation in each word metric type is generally lower within the selected TWGs than among Swedish monosyllabic words.

and Table 1. This question can be answered based on the visual groupings in Figures 5–7. In Figure 5, containing the selected sets of long- and short-vowel TPs, the PD algorithm caused the clustering of the closest vowels within the vowel quadrant. The seemingly close phones [ø̞] and [ɵ] were stopped from being clustered due to their dialectal neutralisation. Similarly, we manually stopped the closely positioned allophones of /ɛ/ and /ø/ from being clustered. Additionally, the consonant groups in Figure 6 – within which all phones differ along the dimension of place of articulation – all consist of articulatorily nearby phones. Thus far, we could have selected the contrasting phones directly from Table 1 and Figure 1. However, when considering the sets of contrasting phones in Figure 7, the clustering is not as apparent from a purely articulatory perspective. For instance, it is not obvious that [l] and [ɳ] are more similar than [l] and [ʝ], or that [m] would be closer to [g] than to [n]. The fact that the model managed to cluster the phones that are clearly closely related (such as those in Figures 5 and 6) indicates that those phones for which the clustering is more ambiguous from an articulatory perspective may also have been appropriately clustered by the computational PD model.

4.1.3 Inclusion of zero-phonemes

We chose to treat empty syllable onsets and codas as being phonemic in nature by adding zero-phonemes in those positions. In doing so, we broadened the definition of minimal variation to include the absence of a phonemic segment. When testing

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

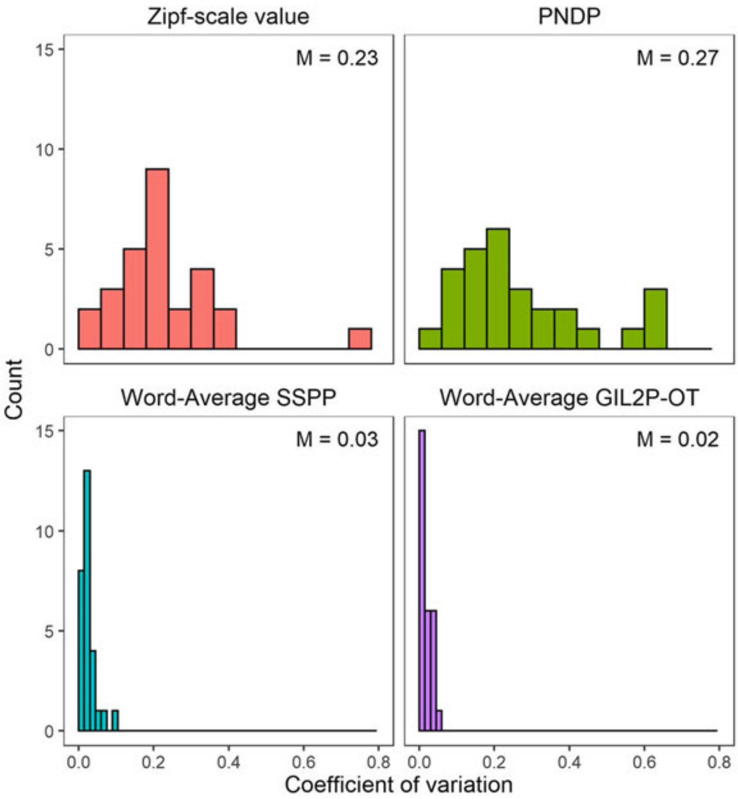

Figure 13. Histograms presenting the distributions of the coefficients of variation within the selected TEST-WORD GROUPS for the four metrics of WORD FREQUENCY (Zipf-scale value), PHONOLOGICAL NEIGHBOURHOOD DENSITY (PNDP), PHONOTACTIC PROBABILITY (word-average SSPP), and ORTHOGRAPHIC TRANSPARENCY (word-average GIL2P-OT). The mean of each distribution is given in the corresponding facet.

the phonemic discrimination ability of people with hearing loss, this is likely useful, as in many cases, certain speech sounds will have fallen below the hearing thresholds at the corresponding frequencies and thus become completely inaudible. Including zero-phonemes in a phonemic discrimination test will help to determine if an incorrect response is due to a misinterpretation of a detected phonetic segment or to a complete lack of phonetic segment detection.

4.2 Construct validity

4.2.1 Accuracy of test-word recordings

The results of the listening experiment, investigating the accuracy of TW recordings for the SiP test, indicated that the majority of the recordings were correctly perceived. However, a few unexpected exceptions occurred. In the case of the word *sjå* [ɧoː∅] 'drudgery', the reasons for the errors could possibly be a word-metric

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

Table 3. Summary of TEST WORDS with repeated errors in the listening experiment.

TEST WORD	TEST-WORD GROUP	Proportion of test sessions containing REPEATED ERRORS (n = 28)		ERRONEOUS RESPONSES
		MALE VOICE	FEMALE VOICE	
pyr [pyːr]	[piːr], [pʉːr], [pyːr]	18%	71%	*pir* [piːr]
sytt [sʏtː]	[sɛ̝tː], [sɪtː], [sʏtː]	0%	14%	*sitt* [sɪtː]
sytt [sʏtː]	[sɪtː], [sʏtː], [sø̞tː]	0%	14%	*sitt* [sɪtː]
sjå [ɧoːØ]	[ØoːØ], [foːØ], [ɧoːØ], [soːØ]	0%	11%	*å* [ØoːØ], *få* [foːØ]
kon [kuːn]	[kuːØ], [kuːn], [kuːɳ]	4%	4%	*korn* [kuːɳ]
tall [talː]	[Øalː], [halː], [palː], [talː]	0%	4%	*pall* [palː]

effect, since the word *sjå* [ɧoːØ] has the lowest WF, PP, and OT values within its group. However, because the errors are only seen with the female voice, there must be other factors involved. One factor could be that the female speaker pronounces [ɧ] in the word *sjå* [ɧoːØ] with very little frication, with the outcome that her [ɧ] is often confused with the softer sounding [f] and the zero-phoneme [Ø]. In our view, pronouncing [ɧ] as a LABIOVELAR APPROXIMANT, rather than a fricative, before a rounded vowel would not be uncommon. However, the pronunciation is obviously causing some difficulty in discriminating between [ɧ], [f] and [Ø], but not between [ɧ] and [s].

The other surprising finding is the apparent difficulty related to identifying both the long and short allophones of the phoneme /y/ occurring in the TWGs [piːr pʉːr pyːr], [sɛ̝tː sɪtː sʏtː], and [sɪtː sʏtː sø̞tː]. For the group with the most erroneous responses ([piːr pʉːr pyːr]), Figures 8–11 show only very minute intragroup differences in WF, PND, and OT. Where differences do exist, they do not disfavour the word containing the /y/. Nevertheless, a possible word-metric effect could stem from PP, for which the words containing the phoneme /y/ have the lowest value in all three groups concerned.

An alternative explanation for the confusion of /y/ for /i/ may be related to our presentation of Lombard recordings in a silent background. The general decrease in spectral tilt and increase of vowel formant frequencies seen in Lombard speech compared to speech in quiet (Van Summers et al. 1988) may have morphed a Lombard /y/ presented in quiet into a /i/. Whether or not this happens with the SiP test materials would have to be determined in a separate study with the SiP test stimuli presented with background sounds.

For the SiP test to have a high level of construct validity, it must actually test the construct of phonemic discrimination as defined by the ability to distinguish between different phonemes. At a minimum, this requires that normal-hearing subjects are able to discriminate between the words in each TWG when their corresponding sound recordings are presented in a silent background, as in the listening experiment in the current study. For the TWGs with repeated incorrect responses described above, this seems not to be the case. Consequently, the

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

construct validity of the concerned TWGs, as a measure of phonemic discrimination, must be questioned. Therefore, the concerned TWGs should possibly be excluded from the final version of the SiP test. Taken together, however, most other TWGs could potentially be used successfully to test the construct of phonemic discrimination.

4.2.2 Top-down and bottom-up processes in speech audiometry

A further critique that threatens the construct validity of the SiP test methodology concerns the fact that discrimination between single words deviates quite a lot from typical real-life speech communication, in which most words occur within a broader context. Notwithstanding, the intent of the SiP test is not primarily to measure someone's ability to utilise top-down processes to integrate the detected speech cues into the broader context of their discourse (see Samuel 2010), but rather to determine the quality of the subject's bottom-up processing of those speech cues. While both types of processes are important in speech perception, hearing rehabilitation interventions (such as hearing aid fittings and cochlear implants) are solely directed toward making the acoustic speech signal more audible to the subject. Hence, it is crucial to be able to reliably quantify any benefits such interventions may have for the bottom-up speech-perception process of individual hearing rehabilitation patients. For the purpose of measuring contextualised speech perception, other tests are available, such as the Swedish version of the Hearing in Noise Test (Hällgren, Larsby & Arlinger 2006).

4.2.3 Word-metric influences

Ultimately, the level of construct validity in the type of closed-set test format selected for the SiP test depends on the extent to which it is only the actual TPs that differ between the presented response alternatives. Since words also tend to differ in other properties influencing lexical access, there would be an imminent risk for reduced construct validity of the SiP test if such biasing confounders had not been controlled. In the current study, we sought to control for such factors by selecting TWGs with a low degree of intragroup variation in Zipf-scale value, PNDP, word-average SSPP, and word-average GIL2P-OT, as defined by their values in the AFC list (Witte & Köbler 2019). We chose these metrics, as we deemed them to be the most important to control for, but we could have selected other metrics as well. Although there is relatively robust support in the literature that factors such as these impact the ease of single-word perception (Pisoni 1996, Luce & Pisoni 1998, Vitevitch & Luce 1998, Ziegler et al. 2003, Dich 2014, Brysbaert et al. 2015, Winkler et al. 2020), the influences upon lexical access from the specific Swedish metrics used in our study have not yet been thoroughly evaluated. However, even if the word metrics would eventually prove to have only an insignificant effect on word perception, the way in which we minimised their variation within the TWGs will hardly have a detrimental impact on the construct validity of the SiP test.

Due to the ceiling effects caused by the optimal listening situation in the listening experiment, the possibility of detecting any major word-metric influences on the test scores is very limited. However, when the SiP test stimuli are eventually presented

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

with disturbing background sounds to people with hearing loss, it will also be possible to determine the influence of the selected word metrics on the SiP test scores. Since the relative degree of remaining word-metric variation within the TWGs was largest for the WF and PND metrics (as indicated by Figure 13), such investigations may show considerably larger influences from those metrics on SiP test scores than from the metrics of PP and OT.

4.2.4 Lombard recordings

As described in the introduction, the final SiP test will present the TWs against an auditory background of naturally occurring disturbing sounds. In such situations, the Lombard effect naturally causes people to increase their vocal effort. Since studies have shown that human speech undergoes several types of temporal and spectral changes in noisy situations that influence its degree of intelligibility (Van Summers et al. 1988), we sought to assimilate such a situation while recording the SiP test TWs by presenting background noise to the speakers. Although this procedure may have caused the SiP test recordings to sound somewhat unnatural in the listening experiment, in which we did not use any background sounds, we believe it will increase the construct validity of the final version of the SiP test.

4.2.5 Multiple test-word recordings

As described above, we made five sound recordings – rather than one, which is usually the case for speech-audiometry tests – for each TW and SiP test voice. Even though this could be regarded as inefficient, having to validate five times as many TW recordings in the current study, the reason behind this choice was to reduce stimulus-specific learning effects in the final version of the SiP test. Although the use of prototype recordings described above likely reduced the variation in pitch, intonation, and intensity patterns across the TW recordings, minor prosodic artefacts are probably unavoidable. We took advantage of this fact by not only minimising the artefacts between different TWs, but also by INTRODUCING such variations within each TW simply by creating multiple recordings of each TW. Shuffling randomly between these, slightly varying, exemplar recordings of each TW in the final version of the SiP test will likely make it harder for patients who take SiP test sessions repeatedly to identify the presented TWs by memorising stimulus-specific artefacts, rather than by phonemic discrimination. In this way, we believe the construct validity of the SiP test will be strengthened.

4.2.6 The use of a previous version of the AFC list

As noted in the methods section above, we based the TW-selection process on an early, unpublished version of the AFC list. The previous version was very similar to the published version of the AFC list, with the exception that vowel lowering of the phonemes /ɛ/ and /ø/ was implemented before all retroflex consonants. This phonological process is correct for monosyllabic words, but not always for bi- or polysyllabic words (Riad 2014, Witte & Köbler 2019). This allophonic change will have affected the PNDP, SSPP, and GIL2P-OT values. The values we used were thus not identical to those in the published AFC list. However, a comparison between the two

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

versions only showed marginal differences. Hence, any consequence for the construct validity of the SiP test stemming from this issue should be very small.

5. Conclusion

We have described the process of developing linguistic materials for a new speech-audiometry method called the SiP test method, as applied to the Swedish language.

The materials developed consisted of 28 different TWGs comprised of real Swedish words, by which a listener's ability to discriminate between phones occurring in monosyllabic Swedish words can be tested. We assembled the TWGs so that the variation of a set of word metrics influencing lexical access was minimised, such that each included TP was contrasted with the most similar other phones in the Swedish language, as determined by a PD measure based on spectrotemporal analysis of sound recordings of contrasting Swedish phones.

Each Swedish SiP test TW was recorded five times by two speakers, one male and one female, and we investigated the accuracy of each TW recording in a listening experiment with normal-hearing subjects. We have made the resulting SiP test speech material available under a Creative Commons Attribution Non-Commercial 4.0 International (CC BY-NC 4.0) licence (https://creativecommons.org/licenses/by-nc/4.0/).

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S0332586521000275

Acknowledgements. This work was supported by Nyckelfonden (grant № OLL-597471) and The Research

Committee (grant #OLL-551401) at Region Örebro County, Sweden. We thank Elin Lundin for her contribution with her voice for the female SiP test recordings. We are also grateful to research audiologists Mathilda Johansson and Sofie Zetterlund at the Audiological Research Centre in Örebro for administering the listening experiment. Finally, we would like to thank four anonymous reviewers, as well as *NJL* editor Gunnar Ólafur Hansson, for very helpful and detailed comments on earlier versions of this article.

Data accessibility statement. The raw data of the current study will be made available upon reasonable request to the Audiological Research Centre at Region Örebro County in Sweden (e-mail: audf@regionorebrolan.se).

Declaration of competing interests. The authors have no conflicts of interests to declare.

Notes

[1] Key to abbreviations: AFC list = The Swedish psycholinguistic database of Witte & Köbler (2019) (see Section 1.4); CV = coefficient of variation (see Section 2.1.2); dB C = decibel C-weighted (see Section 2.2.1); dB FS = decibel full-scale (see Appendix); dB SPL = decibel Sound Pressure Level (see Section 2.2.1); DTW = dynamic time warping (see Appendix); GIL2P-OT = grapheme-initial letter-to-pronunciation orthographic transparency (see Section 1.4); HHIE = Hearing Handicap Inventory for the Elderly (see Section 2.2.2); MVG = minimal-variation group (see Section 1.2); OT = orthographic transparency (see Section 1.4); PB50 = phonemically balanced 50-item word list (see Section 1); PD = phonetic distance (see Section 2.1.1); PND = phonological neighbourhood density (see Section 1.4); PNDP = Zipf-scale weighted phonetic neighbourhood density probability (see Section 1.4); PP = phonotactic probability (see Section 1.4); SD = standard deviation (see Section 2.1.2); SiP = Situated Phoneme (see Section 1); SMA = speech-material annotation (see Appendix); SNR = signal-to-noise ratio (see Section 1.5); SSPP

= normalised stress and syllable-based phonotactic probability (see Section 1.4); TP = test phone (see Section 1.2); TW = test word (see Section 1); TWG = test-word group (see Section 1.6); WF = word frequency (see Section 1.4).

2 These numbers of course depend on the level of detail by which each phonetic segment is described. Here, for example, we have made no distinction between aspirated and unaspirated plosives.

3 We completed the bulk of this study prior to the publication of the AFC list. However, a pre-publication version of the AFC list was available to us for the purpose of this study. The pre-publication version differed from the published version of the AFC list in that a phonological process of vowel lowering of the phonemes /ɛ/ and /ø/ ([ɛː] → [æː], [ɛ̝] → [æ], [øː] → [œː], and [ø̝] → [œ]) was implemented before [r] and some retroflex consonants. This process is correct in monosyllabic words, but is not always in polysyllabic words (Riad 2014). It was therefore not retained in the published version of the AFC list (Witte & Köbler 2019). According to Witte & Köbler (2019), AFC is an abbreviation of the Swedish Audiologiskt forskningscentrum [Audiological Research Centre], which is where the database was developed.

4 Thus, we used the following set of 41 length-reduced phonetic transcriptions: [i ɪ y ʏ ʉ ʊ u e ɛ ɛ̝ ø ø̝ ɵ o ɔ æ œ ɑ a p b t d ʈ ɖ k ɡ f v s ʂ ɕ j ɧ h l r m n ɳ ŋ].

5 When selecting among two or more homophones, we kept the word with the highest average GIL2P-OT. If the GIL2P-OT values were equal between the words, we chose the word with the highest WF.

6 The data in the AFC-list UpperCase field were derived from the corpora studied by Witte & Köbler (2019).

7 When selecting among words with allophonic variation, we included the word with the highest average SSPP. When selecting among words with dialectally neutralised phones, we included the word with the highest WF. If the WF was equal between the words, we selected the word with the highest average SSPP value. When we detected homographs, we selected the word with the highest average SSPP.

8 The microphone settings used while recording were: cardioid polar pattern, high-pass filter of 40 Hz cutoff frequency, no pre-attenuation.

9 The recorded sound was sampled at a rate of 48 kHz and encoded into a 32-bit IEEE floating-point RIFF WAVE file format.

10 We used Sennheiser HD201 headphones since these fitted relatively loosely around the ears and were ventilated enough not to create a strong occlusion effect, which in turn would have counteracted the intended Lombard effect. At the same time, these headphones were isolated enough for the presented noise not to leak into the sound recordings. We calibrated the sound pressure level in these earphones by placing the probe tube of an Otometrics Aurical FreeFit equipment in the concha of each talker, and then adjusting the level of a 1 kHz pure tone to reach approximately 65 dB SPL using the Free-Style function of the FreeFit equipment.

11 We calibrated the loudspeaker using pink noise as the sound source and a condenser microphone placed at the intended position of the participants' heads.

12 Reference number to the decision from the regional ethics review board in Uppsala: dnr 2015/477.

13 As determined by the AFC-list phonetic transcription convention of Witte & Köbler (2019).

14 The word-metric data presented in Section 3 originate from the published version of the AFC list (Witte & Köbler 2019). Since the phones [æ] and [œ] are not used in the AFC-list phonetic transcription convention, we replaced all instances of the phones [æ] and [œ] in the SiP-test material with the corresponding allophones [ɛ̝] and [ø̝] prior to looking up the word-metric values. However, we have retained the phones [æ] and [œ] in the current data presentation.

15 An open-source software library written in Visual Basic.NET that enables reading and writing of wave files with SMA iXML chunk objects is available at https://github.com/witteerik/SHT.Audio.

16 Due to an error in our source code, the first window was counted three times.

References

Agazzi, Birgitta. 2015. *Nyord i svenskan: blogg, fulbryt, pudla, rondellhund och andra nytillskott från A till Ö* [New words in the Swedish language]. Stockholm: Morfem.

ANSI-S3.5. 1997. *American National Standard Methods for Calculation of the Speech Intelligibility Index.* New York: American National Standards Institute.

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

Bashford, James A., Keri R. Riener & Richard M. Warren. 1992. Increasing the intelligibility of speech through multiple phonemic restorations. *Perception and Psychophysics* **51**(3), 211–217. https://doi.org/10.3758/BF03212247.

Bisgaard, Nikolai, Marcel S. M. G. Vlaming & Martin Dahlquist. 2010. Standard audiograms for the IEC 60118-15 measurement procedure. *Trends in Amplification* **14**(2), 113–120. https://doi.org/10.1177/1084713810379609.

Boothroyd, Arthur & Susan Nittrouer. 1988. Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America* **84**(1), 101–114.

Brysbaert, Marc, Michaël Stevens, Pawel Mandera & Emmanuel Keuleers. 2015. The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance* **58**(5), 412–424. https://doi.org/10.1037/xhp0000159.

Carney, Edward & Robert S. Schlauch. 2007. Critical difference table for word recognition testing derived using computer simulation. *Journal of Speech, Language, and Hearing Research* **50**(5), 1203–1209. https://doi.org/10.1044/1092-4388(2007/084).

Clopper, Cynthia G., David B. Pisoni & Adam T. Tierney. 2006. Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology* **17**(5), 331. https://doi.org/10.3766/jaaa.17.5.4.

Deal, Jennifer A., Josh Betz, Kristine Yaffe, Tamara Harris, Elizabeth Purchase-Helzner, Suzanne Satterfield, Sheila Pratt, Nandini Govil, Eleanor M. Simonsick & Frank R. Lin, for the Health ABC Study Group. 2017. Hearing impairment and incident dementia and cognitive decline in older adults: The Health ABC study. *The Journals of Gerontology, Series A: Biological Sciences and Medical Sciences* **72**(5), 703–709. https://doi.org/10.1093/gerona/glw069.

Dich, Nadya. 2014. Orthographic consistency affects spoken word recognition at different grain-sizes. *Journal of Psycholinguistic Research* **43**(2), 141–148. https://doi.org/10.1007/s10936-013-9247-5.

Diehl, Randy L., Andrew J. Lotto & Lori L. Holt. 2004. Speech perception. *Annual Review of Psychology* **55**(1), 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028.

Feeney, Patrick M. & John R. Franks. 1982. Test–retest reliability of a distinctive feature difference test for hearing aid evaluation. *Ear and Hearing* **3**(2), 59–65. https://doi.org/10.1097/00003446-198203000-00002.

Foster, John R. & Mark P. Haggard. 1987. The four alternative auditory feature test (FAAF): Linguistic and psychometric properties of the material with normative data in noise. *British Journal of Audiology* **21**(3), 165–174. https://doi.org/10.3109/03005368709076402.

Friberg, Emilie, Catarina Jansson, Ellenor Mittendorfer-Rutz, Ulf Rosenhall & Kristina Alexanderson. 2012. Sickness absence due to otoaudiological diagnoses and risk of disability pension: A nationwide Swedish prospective cohort study. *PLOS ONE* **7**(1), e29966. https://doi.org/10.1371/journal.pone.0029966.

Ganong, William F., 3rd. 1980. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* **6**(1), 110–125. https://doi.org/10.1037//0096-1523.6.1.110.

Gaskell, Gareth M. & Nicolas Dumay. 2003. Lexical competition and the acquisition of novel words. *Cognition* **89**(2), 105–132. https://doi.org/10.1016/s0010-0277(03)00070-2.

Gelfand, Stanley A. 1998. Optimizing the reliability of speech recognition scores. *Journal of Speech, Language, and Hearing Research* **41**(5), 1088–1102. https://doi.org/10.1044/jslhr.4105.1088.

Gelfand, Stanley A. 2009. *Essentials of Audiology*, 3rd edn. New York: Thieme.

Gold, Ben, Nelson Morgan & Dan Ellis. 2011. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, 2nd edn. Hoboken, NJ: Wiley.

Greenspan, Steven, Raymond Bennett & Ann Syrdal. 1998. An evaluation of the diagnostic rhyme test. *International Journal of Speech Technology* **2**(3), 201–214. https://doi.org/10.1007/BF02111208.

Grunditz, Marie & Lennart Magnusson. 2013. Validation of a speech-in-noise test used for verification of hearing aid fitting. *Hearing, Balance and Communication* **11**(2), 64–71. https://doi.org/10.3109/21695717.2013.782135.

Hagerman, Björn. 1976. Reliability in the determination of speech discrimination. *Scandinavian Audiology* **5**(4), 219–228. https://doi.org/10.3109/01050397609044991.

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

Hällgren, Mathias, Birgitta Larsby & Stig Arlinger. 2006. A Swedish version of the Hearing in Noise Test (HINT) for measurement of speech recognition. *International Journal of Audiology* **45**(4), 227–237. https://doi.org/10.1080/14992020500429583.

Henshaw, Helen & Melanie A. Ferguson. 2013. Efficacy of individual computer-based auditory training for people with hearing loss: A systematic review of the evidence. *PLOS ONE* **8**(5), e62836. https://doi.org/10.1371/journal.pone.0062836.

Hirsh, Ira J., Hallowell Davis, S. Richard Silverman, Elizabeth G. Reynolds, Elizabeth Eldert & Robert W. Benson. 1952. Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders* **17**(3), 321–337. https://doi.org/10.1044/jshd.1703.321.

Holube, Inga & Birger Kollmeier. 1996. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *The Journal of the Acoustical Society of America* **100**(3), 1703–1716. https://doi.org/10.1121/1.417354.

House, Arthur S., Carl E. Williams, Michael H. Hecker & Karl D. Kryter. 1965. Articulation-testing methods: Consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America* **37**, 158–166. https://doi.org/10.1121/1.1909295.

ISO 21388. 2020. *Acoustics – Hearing Aid Fitting Management (HAFM)*. Geneva: International Organization for Standardization.

ISO 7029. 2000. *Acoustics – Statistical Distribution of Hearing Thresholds as a Function of Age*. Geneva: International Organization for Standardization.

Jürgens, Tim & Thomas Brand. 2009. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *The Journal of the Acoustical Society of America* **126**(5), 2635–2648. https://doi.org/10.1121/1.3224721.

Kazanina, Nina, Jeffrey S. Bowers & William Idsardi. 2018. Phonemes: Lexical access and beyond. *Psychonomic Bulletin & Review* **25**(2), 560–585. https://doi.org/10.3758/s13423-017-1362-0.

Kollmeier, Birger, Anna Warzybok, Sabine Hochmuth, Melanie A. Zokoll, Verena Uslar, Thomas Brand & Kirsten C Wagener. 2015. The multilingual matrix test: Principles, applications, and comparison across languages. A review. *International Journal of Audiology* **54** Suppl. 2, 3–16. https://doi.org/10.3109/14992027.2015.1020971.

Kondrak, Grzegorz. 2003. Phonetic alignment and similarity. *Computers and the Humanities* **37**(3), 273–291. https://doi.org/10.1023/A:1025071200644.

Kuk, Francis, Chi-Chuen Lau, Petri Korhonen, Bryan Crose, Heidi Peeters & Denise Keenan. 2010. Development of the ORCA nonsense syllable test. *Ear and Hearing* **31**(6), 779–795. https://doi.org/10.1097/AUD.0b013e3181e97bfb.

Lehiste, Ilse & Gordon E. Peterson. 1959. Linguistic considerations in the study of speech intelligibility. *The Journal of the Acoustical Society of America* **31**(3), 280–286. https://doi.org/10.1121/1.1907713.

Lidén, Gunnar & Gunar Fant. 1954. Swedish word material for speech audiometry and articulation tests. *Acta Oto-Laryngologica. Supplementum* **116**, 189–204. https://doi.org/10.3109/00016485409130295.

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

Martin, Frederick N., Craig A. Champlin & Julie A. Chambers. 1998. Seventh survey of audiometric practices in the United States. *Journal of the American Academy of Audiology* **9**(2), 95–104.

Mielke, Jeff. 2012. A phonetically based metric of sound similarity. *Lingua* **122**(2), 145–163. https://doi.org/10.1016/j.lingua.2011.04.006.

Miller, George A. & Patricia E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America* **27**, 338–352. https://doi.org/10.1121/1.1907526.

Mirman, Daniel, James S. Magnuson, Katharine Graf Estes & James A. Dixon. 2008. The link between statistical segmentation and word learning in adults. *Cognition* **108**(1), 271–280. https://doi.org/10.1016/j.cognition.2008.02.003.

Nakeva von Mentzer, Cecilia, Martina Sundström, Karin Enqvist & Mathias Hällgren. 2018. Assessing speech perception in Swedish school-aged children: Preliminary data on the Listen–Say test. *Logopedics Phoniatrics Vocology* **43**(3), 106–119. https://doi.org/10.1080/14015439.2017.1380076.

Öberg, Marie, Thomas Lunner & Gerhard Andersson. 2007. Psychometric evaluation of hearing specific self-report measures and their associations with psychosocial and demographic variables. *Audiological Medicine* **5**(3), 188–199. https://doi.org/10.1080/16513860701560214.

Oleson, Jacob. J. 2010. Bayesian credible intervals for binomial proportions in a single patient trial. *Statistical Methods in Medical Research* **19**(6), 559–574. https://doi.org/10.1177/0962280209349008.

Olsen, Wayne, Dianne J. Van Tasell & Charles E. Speaks. 1997. Phoneme and word recognition for words in isolation and in sentences. *Ear and Hearing* **18**(3), 175–188. https://doi.org/10.1097/00003446-199706000-00001.

Öster, Anne-Marie. 2006. *Computer-based Speech Therapy Using Visual Feedback with Focus on Children with Profound Hearing Impairments*. Stockholm: KTH Royal Institute of Technology.

Paglialonga, Alessia, Gabriella Tognola & Ferdinando Grandori. 2014. A user-operated test of supra-threshold acuity in noise for adult hearing screening: The SUN (Speech Understanding in Noise) test. *Computers in Biology and Medicine* **52**, 66–72. https://doi.org/10.1016/j.compbiomed.2014.06.012.

Pedersen, Torben Holm & Nick Zacharov. 2018. Sensory evaluation in practice. In Nick Zacharov (ed.), *Sensory Evaluation of Sound*, 61–106. Boca Raton, FL: CRC Press.

Phatak, Sandeep A. & Jont B. Allen. 2007. Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America* **121**(4), 2312. https://doi.org/10.1121/1.2642397.

Pisoni, David. B. 1996. Word identification in noise. *Language and Cognitive Processes* **11**(6), 681–687. https://doi.org/10.1080/016909696387097.

Riad, Tomas. 2014. *The Phonology of Swedish*. Oxford: Oxford University Press.

Rimell, Andrew N., Neil J. Mansfield & Gurmail S. Paddan. 2015. Design of digital filters for frequency weightings (A and C) required for risk assessments of workers exposed to noise. *Industrial Health* **53**(1), 21–27. https://doi.org/10.2486/indhealth.2013-0003.

Risberg, Arne. 1976. Diagnostic rhyme test for speech audiometry with severely hard of hearing and profoundly deaf children. *Speech, Music and Hearing Quarterly Progress and Status Report (STL-QPSR)* 17(2–3). Stockholm: Department for Speech, Music and Hearing, KTH Royal Institute of Technology. http://www.speech.kth.se/qpsr/.

Rødvik, Arne K. 2008. Perception and confusion of speech sounds by adults with a cochlear implant. *Clinical Linguistics & Phonetics* **22**(4–5), 371–378. https://doi.org/10.1080/02699200801919299.

Saffran, Jenny R., Elissa L. Newport & Richard N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language* **35**(4), 606–621. https://doi.org/10.1006/jmla.1996.0032.

Sakoe, Hiroaki & Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055.

Samuel, Arthur G. 2010. Speech perception. *Annual Review of Psychology* **62**(1), 49–72. https://doi.org/10.1146/annurev.psych.121208.131643.

Shadish, William R., Thomas D. Cook & Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Sigurd, Bengt. 1965. *Phonotactic Structures in Swedish*. Ph.D. dissertation, Lund University.

Smeds, Karolina, Florian Wolters & Martin Rung. 2015. Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of the American Academy of Audiology* **26**(2), 183–196. https://doi.org/10.3766/jaaa.26.2.7.

Sommers, Mitchell S., Karen Iler Kirk & David B. Pisoni. 1997. Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners, I: The effects of response format. *Ear and Hearing* **18**(2), 89–99. https://doi.org/10.1097/00003446-199704000-00001.

Statistics-Sweden. 2019. Living conditions surveys (ULF/SILC): Table on health – disabilities 2018. https://www.scb.se/hitta-statistik/statistik-efter-amne/levnadsforhallanden/levnadsforhallanden/undersokningarna-av-levnadsforhallanden-ulf-silc/pong/tabell-och-diagram/halsa/halsa–fler-indikatorer/.

Streiner, David L. & Geoffrey R. Norman. 2008. *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford: Oxford University Press.

Sumby, William H. & Irwin Pollack. 1954. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America* **26**(2), 212–215. https://doi.org/10.1121/1.1907309.

Thornton, Aaron R. & Michael J. M. Raffin. 1978. Speech-discrimination scores modeled as a binomial variable. *Journal of Speech, Language, and Hearing Research* **21**(3), 507–518. https://doi.org/10.1044/jshr.2103.507.

van Heuven, Walter J. B., Pawel Mandera, Emmanuel Keuleers & Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* **67**(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521.

Van Summers, Walter, David B. Pisoni, Robert H. Bernacki, Robert I. Pedlow & Michael A. Stokes. 1988. Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America* **84**(3), 917–928. https://doi.org/10.1121/1.396660.

Vitevitch, Michael S. & Paul A. Luce. 1998. When words compete: Levels of processing in perception of spoken words. *Psychological Science* **9**(4), 325–329. https://doi.org/10.1111/1467-9280.00064.

Vitevitch, Michael S. & Paul A. Luce. 2004. A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods Instruments & Computers* **36**(3), 481–487. https://doi.org/10.3758/Bf03195594.

Välimaa, Taina T., Taisto K. Määttä, Heikki J. Löppönen & Martti J. Sorri. 2002. Phoneme recognition and confusions with multichannel cochlear implants: Consonants. *Journal of Speech, Language, and Hearing Research* **45**(5), 1055–1069. https://doi.org/10.1044/1092-4388(2002/085).

Wagener, Kirsten C., Martin Hansen & Carl Ludvigsen. 2008. Recording and classification of the acoustic environment of hearing aid users. *Journal of the American Academy of Audiology* **19**(4), 348. https://doi.org/10.3766/jaaa.19.4.7.

Wichmann, Felix & Nichola J. Hill. 2001. The psychometric function, I: Fitting, sampling, and goodness of fit. *Perception and Psychophysics* **63**(8), 1293–1313. https://doi.org/10.3758/BF03194544.

Winkler, Alexandra, Rebecca Carroll & Inga Holube. 2020. Impact of lexical parameters and audibility on the recognition of the Freiburg Monosyllabic Speech Test. *Ear and Hearing* **41**(1), 136–142. https://doi.org/10.1097/AUD.0000000000000737.

Witte, Erik & Susanne Köbler. 2019. Linguistic materials and metrics for the creation of well-controlled Swedish speech perception tests. *Journal of Speech, Language, and Hearing Research* **62**(7), 2280–2294. https://doi.org/10.1044/2019_JSLHR-S-18-0454.

Woods, David L., Tanya Arbogast, Zoe Doss, Masood Younus, Timothy J. Herron & E. William Yund. 2015. Aided and unaided speech perception by older hearing impaired listeners. *PLOS ONE* **10**(3). https://doi.org/10.1371/journal.pone.0114922.

World Medical Association. 2013. World medical association declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA – Journal of the American Medical Association* **310**(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053.

Yu, Tzu-Ling J. & Robert S. Schlauch. 2019. Diagnostic precision of open-set versus closed-set word recognition testing. *Journal of Speech, Language, and Hearing Research* **62**(6), 2035–2047. https://doi.org/10.1044/2019_jslhr-h-18-0317.

Ziegler, Johannes C., Mathilde Muneaux & Jonathan Grainger. 2003. Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language* **48**(4), 779–793. https://doi.org/10.1016/S0749-596X(03)00006-8.

Appendix. Calculation of phonetic distance between Swedish phones

This appendix outlines the methods used to calculate PHONETIC DISTANCE (PD) between Swedish phones used in the current study. We based the PD calculations on sound recordings of a set of real Swedish monosyllabic words that altogether formed 106 word groups, within which each member formed minimal pairs with all other members. We refer to this type of word group as a MINIMAL-VARIATION GROUP (MVG). We formed the MVGs using the same techniques described in the main article for the selection of candidate TWGs for the SiP test, employing the same earlier version of the AFC list (Witte & Köbler 2019).

To minimise the number of MVGs while including as many phonemic contrasts as possible, we went through all generated MVGs, starting by including the largest available group and subsequently including smaller groups, but only if they contributed new phonemic contrasts to the already included MVGs. In this way, we identified 106 MVGs contrasting all phones that occur contrastively in Swedish monosyllabic words. Supplement 1 contains these groups.

We then recorded all words occurring in the selected MVGs in digital sound files using the voice of the first author. The recordings were made in a sound-treated booth using a Neumann TLM 103 condenser

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press

microphone. The sample rate used was 48 kHz, and the data were stored as 32-bit IEEE floating-point numbers in RIFF WAVE format files. To extract the sound file sections containing the phones of interest, we manually determined the location of all phonetic segment boundaries occurring in the recordings and stored them within the SPEECH-MATERIAL ANNOTATION (SMA) object in an iXML chunk (www.ixml.info) of each sound file. The SMA specification, which we developed in this study, can store linguistic segmentation data along with sound level measurements directly in the sound file. Supplement 8 provides a detailed specification of the SMA iXML object.[15] To approximately equalise the loudness of each sound file, we normalised their waveforms to a C-weighted maximum level of −23 dB FS using an integration time of 0.1 seconds. We accomplished C-weighting with sixth-order IIR FILTERING using the coefficients presented in Rimell, Mansfield & Paddan (2015). We have made these sound recordings available in Supplement 2 under the Creative Commons Attribution Non-Commercial 4.0 International (CC BY 4.0) licence (https://creativecommons.org/licenses/by/4.0/). Then, based on the extracted phone sections of these sound files, we calculated the PD of each phone to all other phones (to which it was contrasted in the material) using the following SPECTROTEMPORAL analysis method.

For each pair of contrasting phones, we first computed a time series expressing the BARK SPECTRA within 95% overlapping HAMMING-filtered, 0.1 seconds long, TIME-DOMAIN WINDOWS, across the entire duration of each phone. For each time-domain window, we derived the corresponding Bark spectrum by first calculating an 8,192-point FAST FOURIER TRANSFORM of the particular time-domain window, and then averaging the resulting spectral amplitudes within 234 triangular FREQUENCY-DOMAIN windows, each 2 Bark wide and with 90% overlap.[16] The upper panes of Figure A1 depict examples of such Bark spectra for the phones [iː] and [oː] (derived from the words *rid* [riːd] 'ride' and *råd* [roːd] 'advice'). The centre frequencies of these frequency-domain windows ranged from 100 Hz to 17,242 Hz. We then used DYNAMIC TIME WARPING (DTW) to identify the most efficient path through a matrix of EUCLIDEAN DISTANCES between the spectral representations stored in the time-domain windows corresponding to the two phones compared. We utilised the symmetric 0-algorithm of Sakoe & Chiba (1978) for this DTW algorithm, with the types of global constraints described by Gold, Morgan & Ellis (2011:343). Thus, the DTW SEARCH SUBSPACE was limited to the region specified in Figure A2. In addition, local constraints, only allowing the DTW to make one time warp at a time in the same warp direction, prevented the selected path through the DTW matrix from making sharp turns within the search subspace. The bottom left panel of Figure A1 visualises the distance matrix generated in the comparison of the phones [iː] and [oː], along with the most efficient path (plotted as a black line). We then summed the distance values along the path selected by the DTW, as seen in the bottom right panel of Figure A1, and then normalised them by the sum of the number of time-domain windows in the compared sounds (see Sakoe & Chiba 1978). Finally, to ensure that speech sounds of high spectral similarity but different temporal durations would still be graded as different, we multiplied the normalised DTW output by the ratio of the duration of the longer sound to that of the shorter. The attained value comprised our measure of PD. As such, we defined PD on an arbitrary unit scale with a true zero point. Supplement 3 shows the resulting PDs between all phonemically contrasting Swedish phones.

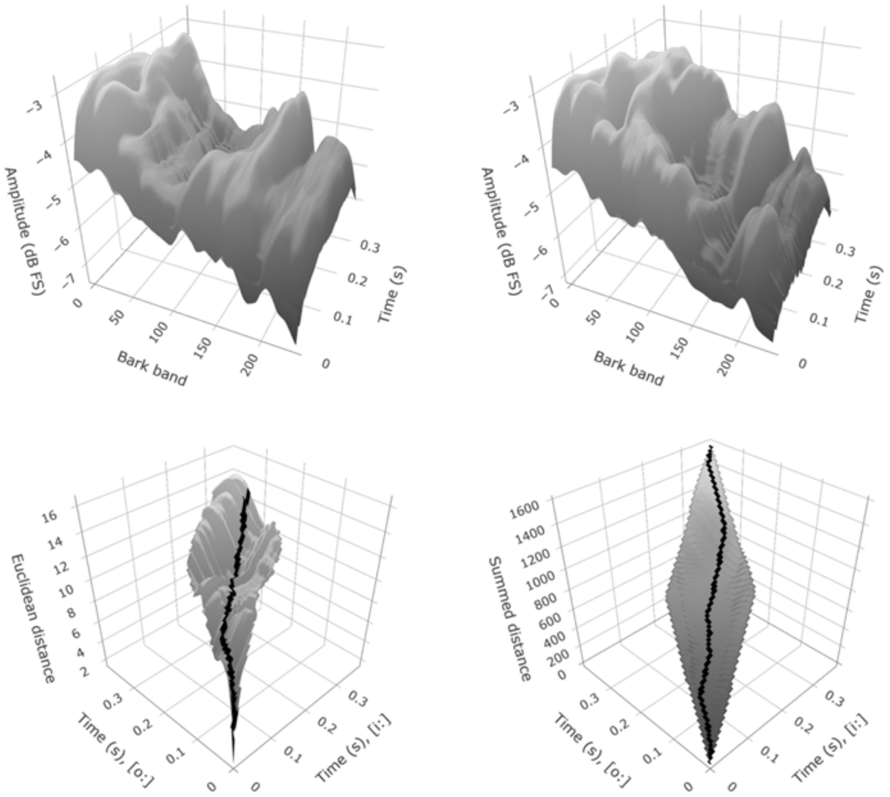

Figure A1. Visualisations of three steps in the calculation of phonetic distance (PD). The top panels display Bark spectra for the phone [iː] in the word *rid* [riːd] 'ride' (left pane) and for the phone [oː] in the word *råd* [roːd] 'advice' (right pane). The bottom left panel depicts a matrix of Euclidean distances between different time-domain windows in the [iː] and [oː] Bark spectra. The bottom right panel shows the summation of Euclidean distance values along the path selected by a dynamic time warping (DTW) algorithm. The selected path is marked as a black line in both lower panes.

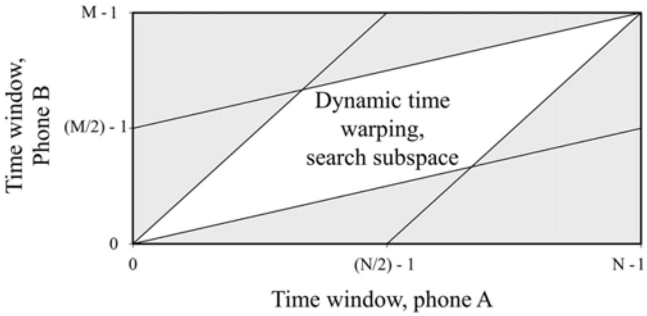

Figure A2. The DYNAMIC TIME WARPING (DTW) subspace (white area) used when comparing the frequency-domain content of different time-domain windows of phones A and B. N and M represent the number of time-domain windows in phone A and phone B, respectively.

Cite this article: Witte E, Ekeroot J, and Köbler S (2024). The development of linguistic stimuli for the Swedish Situated Phoneme test. *Nordic Journal of Linguistics* **47**, 73–110. https://doi.org/10.1017/S0332586521000275

https://doi.org/10.1017/S0332586521000275 Published online by Cambridge University Press