Quantifying the intelligibility of speech in noise for non-native listeners

Sander J. van Wijngaarden,[a] Herman J. M. Steeneken, and Tammo Houtgast
TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands

(Received 9 July 2001; accepted for publication 9 January 2002)

When listening to languages learned at a later age, speech intelligibility is generally lower than when listening to one's native language. The main purpose of this study is to quantify speech intelligibility in noise for specific populations of non-native listeners, only broadly addressing the underlying perceptual and linguistic processing. An easy method is sought to extend these quantitative findings to other listener populations. Dutch subjects listening to Germans and English speech, ranging from reasonable to excellent proficiency in these languages, were found to require a 1–7 dB better speech-to-noise ratio to obtain 50% sentence intelligibility than native listeners. Also, the psychometric function for sentence recognition in noise was found to be shallower for non-native than for native listeners (worst-case slope around the 50% point of 7.5%/dB, compared to 12.6%/dB for native listeners). Differences between native and non-native speech intelligibility are largely predicted by linguistic entropy estimates as derived from a letter guessing task. Less effective use of context effects (especially semantic redundancy) explains the reduced speech intelligibility for non-native listeners. While measuring speech intelligibility for many different populations of listeners (languages, linguistic experience) may be prohibitively time consuming, obtaining predictions of non-native intelligibility from linguistic entropy may help to extend the results of this study to other listener populations.
[DOI: 10.1121/1.1456928]

PACS numbers: 43.71.Gv, 43.71.Hw [CWT]

I. INTRODUCTION

Most people know from personal experience that ''non-native'' speech communication is generally less effective than purely ''native'' speech communication. This is readily verified by listening to foreign-accented speech in one's own language, or by trying to comprehend speech in a foreign language that is not fully mastered. It is also known that the intelligibility of speech depends strongly on the experience with the target language by listeners as well as talkers (e.g., Flege, 1992; Strange, 1995). Especially under adverse conditions (noise, reverberation, background babble), non-native speech communication tends to be less effective (Lane, 1963; Gat and Keith, 1978; Mayo *et al.*, 1997; Nábělek and Donahue, 1984).

Non-native speech has been studied extensively, from the perspective of production as well as perception. Usually, the objective of second-language (L2) speech studies is to contribute to a more profound insight into the complicated processes underlying speech perception. By contrast, our approach starts out by studying the intelligibility effect of non-nativeness in its own right. This information, when properly quantified, is expected to be directly applicable in more engineering-oriented disciplines associated with speech communication (speech intelligibility in room acoustics, design of communication systems). Our findings are also intended to be used for incorporating ''the non-native factor'' in existing speech intelligibility prediction models, such as the speech transmission index (STI; Steeneken and Houtgast, 1999) and the speech recognition sensitivity model (SRS; Müsch and Buus, 2001). They may also be useful in the field of clinical audiology, where the effects of hearing loss on speech intelligibility may be confounded with the effects of being raised in a ''foreign'' language.

In this study, the focus will be on the intelligibility effects of non-nativeness from the perspective of speech perception only: we will try to quantify the extent to which a population of L2 learners will suffer reduction of speech intelligibility when listening to a second language.

A great number of variables will influence the speech understanding process for a certain population of non-native listeners. First of all, the relation between the native language and the target (second) language is of importance. Between languages that are relatively similar (in terms of functional phonetic contrasts, phonology, etc.) different effects may be observed than between languages that have very little in common. As already stated above, an important factor is also the population's average experience with the second language (number of years since the language was first learned, intensity of use). Age of acquisition of the second language is another important variable (Flege, 1995; Flege *et al.*, 1997; Mayo *et al.*, 1997), as well as the amount of continued native language use (Meador, 2000). In order to be able to predict the size of any intelligibility effect involving non-native listeners, the population of listeners should be specified in terms of (at least) these factors.

Various studies have produced quantitative results of non-native speech intelligibility for specific subject populations. Florentine *et al.* (1984), for example, reported reduced speech intelligibility in noise for non-native subjects. The

[a]Electronic mail: vanwijngaarden@tm.tno.nl

 0001-4966/2002/111(4)/1906/11/$19.00

speech-to-noise ratio required for 50% intelligibility of redundant sentences was 4 to 15 dB higher for French learners of the English language than for native English listeners, depending on experience. Florentine (1985) also found that non-native listeners were less able to take advantage of context; the difference between natives and non-natives was smaller for low-predictability sentences than for high-predictability sentences. These findings are supported, for instance, by the experiments of Mayo *et al.* (1997). This is contrary to predictions by Koster (1987), who conducted a series of linguistic experiments with Dutch subjects who were studying to become English teachers. By systematically varying the predictability of a test word through manipulation of its context, he found that the effect of semantic constraints on word recognition was of the same magnitude for native and non-native listeners. A closer investigation of the use of contextual information by non-native listeners is therefore needed.

Experiments concerning non-native speech intelligibility in noise will be described in Sec. II of this article: speech reception threshold (SRT) results are presented, which will allow a broad quantitative comparison between native and non-native speech intelligibility in noise. In Sec. III, this comparison will be refined by looking at the slope of the psychometric function in a sentence recognition task. In Sec. IV we will describe experiments exploring the relation between non-native sentence recognition and redundancy-related measures.

II. INTELLIGIBILITY THRESHOLD OF SPEECH IN NOISE FOR NON-NATIVE LISTENERS

A. Method

An interesting topic in relation to non-native speech perception is the use of word context. This means that speech intelligibility for non-native listeners is best measured using longer phrases (sentences). For measuring sentence intelligibility under the influence of noise, several proven methods are available, among which is the speech reception threshold (SRT; Plomp and Mimpen, 1979). The SRT method, used for all intelligibility experiments described in this article, is an adaptive method that measures the speech-to-noise ratio at which 50% of the tested sentences are perceived correctly. All listeners were Dutch; SRT tests were carried out with the same group of listeners, using sentences in three different languages: Dutch (D), English (E), and German (G).

1. Subjects

In order to allow meaningful interpretation of the intelligibility results obtained through SRT experiments, a well-defined population of test subjects has to be chosen. Mean scores across subjects will only be meaningful if the group of subjects is homogeneous in terms of L2 proficiency, age, level of education, and other factors influencing second-language skills.

Two main groups of subjects were recruited for this experiment. Group I was recruited following fairly strict guidelines. The recruiter used a ''checklist'' to make sure that only subjects were accepted that matched a set of predefined criteria. Group I consisted of nine tri-lingual Dutch university students of various disciplines (not including languages or phonetics), aged 18–24 years, who considered English their second language and German their third language. All had first learned both English and German, written and orally, during secondary education (Dutch high school), all starting with English at age 12 or 13, and with German at age 13 or 14. For each individual subject, the self-reported overall proficiency (rated on a 5-point scale) was higher for English (mean rating 3.7) than for German (mean rating 2.9). All individual subjects had a much more frequent use of English than of German: all reported daily use of English (reading and/or listening), while use of the German language was typically weekly to monthly.

Subject group II, consisting of 11 subjects, was matched to group I in terms of age (18–24) and level of education, but without the strict requirements on experience with English and German. Group II subjects were only required to be able to understand spoken and written English and German above a certain minimum level. The spread in German proficiency was therefore larger (mean rating 3.3); the frequency of use of the German language varied from daily to yearly for group II. For English, mean self-reported proficiency and frequency of use of group II turned out to be just as good as of group I (mean rating 3.4). This is probably due to demographic and educational causes: Dutch university students are generally quite proficient in English. The fact that young Dutch people mainly watch English-spoken television with Dutch subtitles may also be part of the explanation.

In addition to the main subject groups I and II, two control groups were recruited: three native German and three American subjects. These control groups were used to verify that the implementation of the SRT test (sentence material and talkers) was equivalent across languages.

2. Procedure

The SRT test gives a robust measure for sentence intelligibility in noise, corresponding to the speech-to-noise ratio that gives 50% correct response of short redundant sentences. In the SRT testing procedure, masking noise is added to test sentences in order to obtain speech at a known speech-to-noise ratio. The standard masking noise spectrum (as applied in the experiments described in this article) is equal to the long-term average spectrum of the test sentences. After presentation of each sentence, the subject responds by orally repeating the sentence to an experimenter. The experimenter compares the response with the actual sentence. If every word in the responded sentence is correct, the noise level for the next sentence is increased by 2 dB; after an incorrect response, the noise level is decreased by 2 dB. The first sentence of a list of 13 sentences is repeated until it is responded correctly, using 4-dB steps. This is done to quickly converge to the 50% intelligibility threshold. By taking the average speech-to-noise ratio over the last ten sentences, the 50% sentence intelligibility threshold (SRT) is obtained.

During the experiments, the subjects (listeners) were seated in a sufficiently silent room. A set of Sony MDR-CD770 headphones was used to present the recorded sentences diotically to the listeners. All subjects participated in a brief training session before taking part in the actual experi-

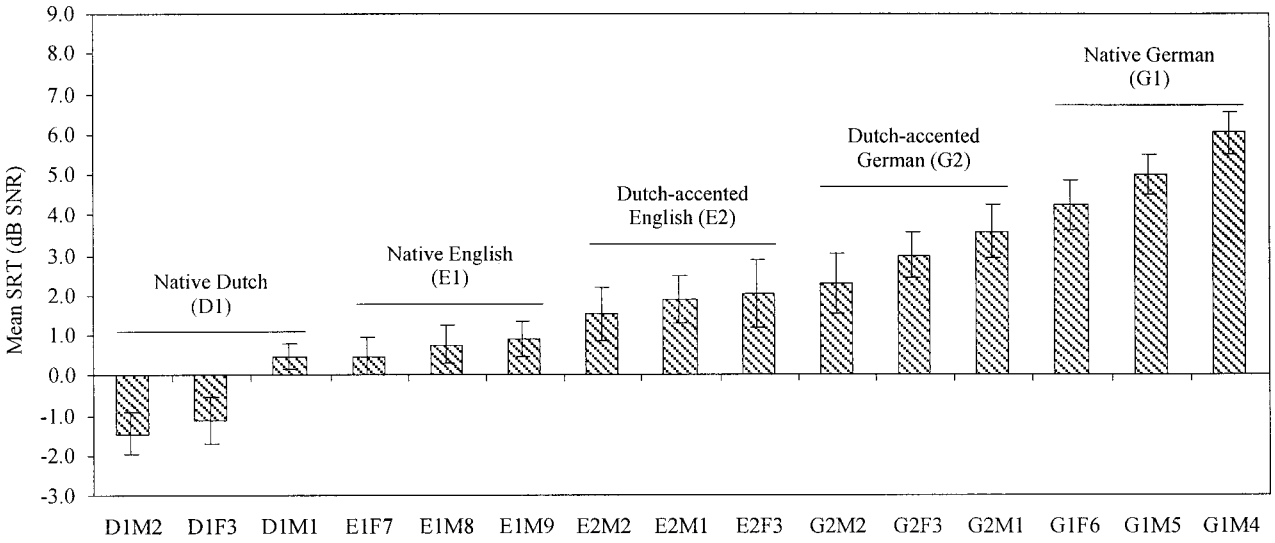

FIG. 1. Mean SRT results of subject group I per individual talker ($N=9$). All listeners were Dutch students, speaking English as a second language and German as a third language. Speech material was in Dutch (D1), English (E1 and E2), and German (G1 and G2). Non-native talkers (E2 and G2) were all Dutch. The talkers are labeled according to language group (e.g., D1), gender (M for male, F for female), and a unique number (1–9) for each talker. The error bars represent the standard error.

ments. None of the subjects had heard or read the test sentences before the experiments; each sentence was used only once with each subject, to avoid memory effects.

3. Stimuli

In order to be able to carry out speech intelligibility tests, suitable speech material has to be collected. A set of 130 standardized Dutch SRT sentences (10 lists of 13 sentences) were ''translated'' to German and English by native talkers of these languages with phonetic expertise and experience in speech research. This ''translation'' did not perfectly preserve the literal meaning of the sentences; the aim was to obtain the same context, complexity and length (number of syllables) in all languages. A procedure for obtaining multi-lingual speech databases for SRT tests, which gives equivalent results across languages, was described by van Wijngaarden *et al.* (2001). The sentences were recorded as spoken by native talkers of Dutch, German, and American English (referred to from hereon as D1, G1, and E1). Additionally, Dutch talkers (the same talkers as for the D1 experiment) also recorded English and German sentences (G2 and E2). Recordings were made for a total of nine talkers: three for each native language (two male, one female); because of the fact that the Dutch talkers recorded three sets of sentences (D1, G2, and E2), a total of 15 sets of recorded sentences was collected.

Talkers did not demonstrate any speaking disorders, and were informally estimated to have more or less average clarity of articulation. Influences of regional accents (deviations from the preferred pronunciation in the respective languages), when noticeable at all, were minor.

B. Results

1. Fully native baseline SRT scores

Conclusions regarding the effects of non-nativeness can only be drawn if the SRT implementation that is used is also independent of language. In other words, we need to make sure that the precautions taken in the ''translation'' of the test sentences were effective in making the German and English test equal to the original Dutch test. This was verified by conducting ''fully native'' SRT tests in all three languages (three talkers per language; three English listeners, three German listeners, and 20 Dutch listeners).

The mean SRT was close to −1 dB in all of the languages (−0.8 for Dutch, −1.0 for English, and −1.1 for German). None of the differences in native SRT is statistically significant. This indicates that the performance of the SRT test is language independent.

Compared to SRT results found with thoroughly optimized SRT databases, a mean SRT of −1 dB may seem high. For a nonoptimized SRT test in Dutch (but with specifically selected talkers, which his not the case in the multi-lingual SRT test), Versfeld *et al.* (2000) report a mean SRT of −1.8 dB. The difference can most likely be attributed to the concessions done to keep the recording procedure practical, and the absence of a strict talker selection regime [see van Wijngaarden *et al.* (2001), for more details].

2. SRT scores of group I

Group I, the homogeneous group of nine trilingual Dutch subjects, participated in a SRT experiment in which subjects were presented with Dutch, German, and English speech. In addition to the SRT sentences by (native) G1 and E1 talkers, they were also presented with speech by the three Dutch talkers in German and English (G2 and E2). In this latter case, the overall intelligibility will not only be affected by non-native speech perception, but also by non-native speech production. The results from this experiment, separated by individual talker, are given in Fig. 1.

The talkers in Fig. 1 are grouped by language, and rank ordered according to mean SRT for all nine listeners. The effect of non-native perception of English (difference between D1 and E1 scores) is relatively small; the mean differ-

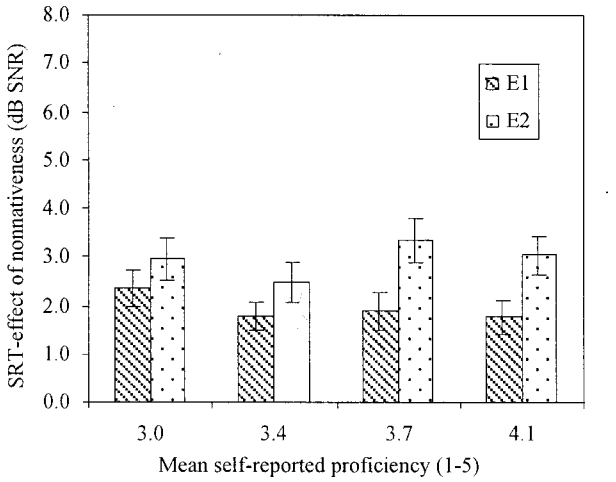

FIG. 2. The effect of non-nativeness (difference between native and non-native SRT) for subgroups of five subjects differing in self-reported proficiency. The non-native language is English. The error bars indicate the standard error (five subjects, three speakers; $N=15$).

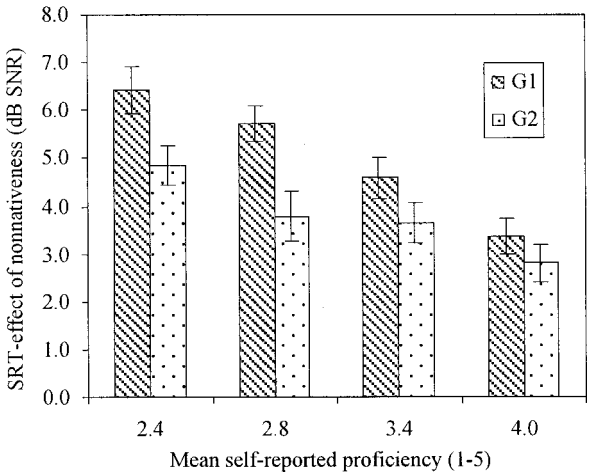

FIG. 3. The effect of non-nativeness (difference between native and non-native SRT) for subgroups of five subjects differing in self-reported proficiency. The non-native language is German. The error bars indicate the standard error (five subjects, three speakers; $N=15$).

ence in SRT is 1.4 dB. The mean difference between D1 and G1 is much larger: 5.8 dB. Different deficits for English and German were to be expected; the difference in proficiency and intensity of use have a clear effect on intelligibility. Compared to earlier results from similar studies in other languages (e.g., Buus *et al.*, 1986; Mayo *et al.*, 1997), the G1 deficit matches expectations, but the E1 deficit is smaller than expected for late bilinguals. The frequent ''early'' exposure of young Dutch people to English speech on television may be part of the explanation.

It is interesting to compare the scores for E1 (American English talkers) and E2 (Dutch talkers of the English language). The Dutch listeners do not benefit from hearing their ''own'' non-native accent in a second language: the native English talkers provide a better intelligibility. This is consistent with earlier findings by van Wijngaarden (2001) for the reverse situation (American subjects listening to Dutch sentences). For G1 and G2, the effect is exactly opposite: the Dutch listeners do experience better intelligibility in German if the talkers have a Dutch accent.

3. SRT scores of groups I and II together (group I+II)

The same SRT conditions presented to group I were also tested with group II. By combining the data of groups I and II, analysis based on a larger group of 20 subjects (which we will call ''group I+II'') may be carried out, which will be more diverse in terms of their proficiency, at least in German. This allows us to study the effect of proficiency and experience on speech intelligibility.

In Figs. 2 and 3, combined SRT results for group I+II are given. Scores for the 20 subjects were divided into four subgroups of five subjects, according to the self-reported proficiency of the subjects. The leftmost subgroup in each figure is the subgroup with the lowest self-reported proficiency, the rightmost is the one with the highest proficiency. Although Fig. 2 (English) and Fig. 3 (German) are based on scores of the same 20 subjects, the division into subgroups is different. The division enables investigation of the effect of proficiency on intelligibility. This is not easily done on the basis of individual proficiency ratings, since these tend to be fairly unreliable.

The results of Figs. 2 and 3 are not simply mean SRT scores on the German and English sentences, but rather the difference of these scores with the scores on the Dutch sentences. This difference is a direct measure of the effect of non-nativeness on speech intelligibility. By taking this difference, a correction is also applied for small differences in (native) Dutch SRT scores between the subgroups.

Figure 2 shows no significant effects of self-reported proficiency. All subjects (also from group II) showed a good command of the English language.

Whereas Fig. 2 does not show any systematic relation between intelligibility and self-reported proficiency, Fig. 3 demonstrates that such a relation can exist. For authentic, unaccented German speech, the intelligibility is higher (the effect of non-nativeness smaller) to the subgroups with higher proficiency ratings. The most proficient subgroup, for example, shows a significantly smaller effect ($p<0.05$) than all of the other three subgroups for G1 talkers. With the exception of the differences between neighboring subgroups, all other differences for G1 talker in Fig. 3 are also statistically significant ($p<0.05$; *t*-tests used to compare the means between subgroups).

The scores for G2 talkers (Dutch-accented German speech) appear to show the same trend. Here, however, the only difference between subgroups that is statistically significant is the difference between the least proficient and the most proficient subgroup ($p<0.01$).

According to Fig. 2, E1 speech (authentic American English pronunciation) tends to be somewhat more intelligible to non-native Dutch listeners than (accented) English speech by Dutch talkers. This same effect was observed in Fig. 1, and appears to be relatively independent of (small) differences in proficiency.

Figure 3 shows, much the same as Fig. 1, a difference between G1 and G2 intelligibility that is contrary to the difference between E1 and E2. The difference between G1 and

G2 intelligibility appears to decrease with proficiency. The two subgroups with the lower self-reported proficiency differ significantly between G1 and G2; the differences are not significant for the other two (more proficient) subgroups.

It is clear that even subjects that give themselves high ratings for German proficiency have more problems understanding spoken German than the average subject has understanding spoken English. This is observed by comparing the effect of non-nativeness of the most proficient (rightmost) subgroup in Fig. 3 (German) to the least proficient (leftmost) subgroup in Fig. 2 (English); the performance in English appears to be still better than in German, although it is difficult to establish clear statistical proof for this.

Please note that the mean proficiency ratings for the subgroups are only used as relative rankings of proficiency to obtain a division into subgroups. These ratings hold no absolute value; the ratings for English may, for instance, not be directly compared to the ratings for German. The reason for this is that the subjects tend to rate themselves in relation to the performance of their peer group. A more objective measure of proficiency is needed to understand how the results reported in Fig. 3 are related to the results in Fig. 2 (in other words, how the differences in effects between English and German are explained in terms of differences in proficiency). This will be further explored in Sec. IV.

III. STEEPNESS OF THE PSYCHOMETRIC FUNCTION FOR NON-NATIVE SENTENCE INTELLIGIBILITY

A. Methods

The SRT results given in Sec. II characterize the psychometric function of sentence intelligibility by a single value: the SNR for which 50% sentence recognition occurs. However, much speech communication in real life takes place at speech-to-noise ratios corresponding to other levels of sentence intelligibility than 50%. We would therefore like to know the full psychometric function, so that we can predict the SNR necessary to meet any intelligibility criterion. This is especially relevant since the slope of the psychometric function is known to differ between native and non-native listeners (e.g., Mayo *et al.*, 1997).

The straightforward way of obtaining a full psychometric function is by sampling the curve at a fixed set of speech-to-noise ratios. This can be a rather laborious process. There is a theoretical possibility to extract additional information about the psychometric function from standard SRT measurements (Plomp and Mimpen, 1979). Unfortunately, the SRT experiments underlying Figs. 1 and 2 do not include enough individual subject responses at various SNR values to allow an accurate estimate of the steepness of the psychometric function.

A compromise between sampling the entire psychometric function and estimation of the steepness from standard SRT tests was chosen: first the standard SRT was measured, then the percentage of correctly responded sentences was measured directly at four speech-to-noise ratios around the SRT. Next, the psychometric function was fit through these points.

Slopes of the psychometric function will be compared across languages. In a fully native setting (talker and listener), the SRT in Dutch, English, and German was found to be equal, leading to the conclusion that SRT results can be compared across languages in a straightforward way. For the slope of the psychometric function, this firm baseline was not established, but there are no reasons to expected considerable differences.

1. Subjects, stimuli, and conditions

A new group of 15 trilingual subjects was recruited, matching subject group I (nine subjects) on all relevant parameters. Since SRT subjects must be unacquainted with the sentence material, and the available material was limited to ten lists per language, the subjects from experiment I could not participate in this experiment. For the same reason, the conditions tested in this experiment do not include all talkers from experiment I. The three (baseline) Dutch talkers were included, as well as talker E1M8 (see Fig. 1) to represent the English talkers and talkers G1M5 to represent the German talkers. Dutch talker No. 3 was also included as an L2 talker of German (labeled G2F3 in Fig. 1) and English (E2F3). Material of each talker was presented to 5 subjects out of the group of 15.

2. Procedure

First of all, a standard SRT test was carried out for each subject in each condition. Next, the percentage of correctly repeated sentences was determined at SNR values differing by -4, -2, $+2$, and $+4$ dB relative to the SRT. The same criterion was used as in a standard SRT test: the subjects had to be able to correctly repeat the entire sentence for the presentation to be considered ''correct.'' At each SNR value, a single list of SRT sentences (13 sentences) was presented.

Following this procedure, five points of the psychometric function were obtained (including the SRT at 50%) per subject per condition. A cumulative normal distribution was fit through these points using a nonlinear least-squares approach (Gauss–Newton method). Hence, the model assumed for the psychometric function was a cumulative normal distribution. Effectively, two parameters of the distribution were fit: the mean and the standard deviation. The mean of the distribution corresponds to the SRT, while the steepness of the psychometric function at 50% intelligibility is directed related to the standard deviation (Versfeld *et al.*, 2000). The steeper the psychometric function, the stronger the effect of a difference in speech-to-noise ratio on speech intelligibility.

B. Results

The speech reception threshold and the distribution mean obtained by fitting the psychometric function through observation data are essentially different estimates of the same variable: the 50% point of the psychometric function. Both estimates were found to yield very similar results.

The estimated slopes of the psychometric function around 50% intelligibility are given in Fig. 4.

Even at first sight, the steepness of the psychometric function clearly has an inverse relation with the SNR at the 50% point: talkers with higher values of the SRT (50% point)

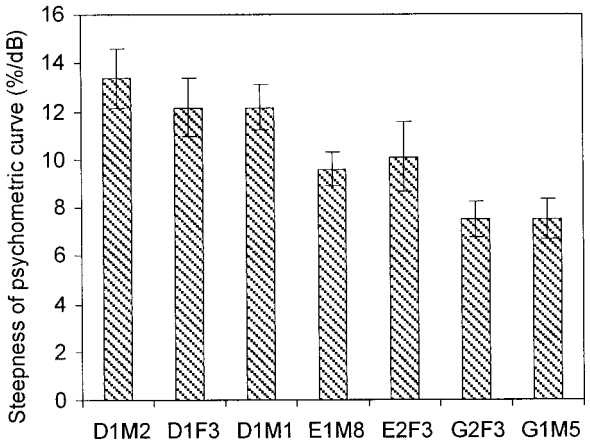

FIG. 4. Estimates of the steepness (slope at the 50% point) of the psychometric function for seven individual talkers. Error bars indicate the standard error of the estimates (five subjects; $N=5$).

have lower steepness, while language appears to be the explaining variable. The statistical significance of the differences in Fig. 4 was investigated by means of a Newman–Keuls test, after finding a significant effect in a one-way ANOVA. None of the differences between talkers speaking the same language was significant. The difference between G2F3 and E2F3, as well as the difference between E1M8 and D1M1, is also not significant. All other differences in Fig. 4 are statistically significant ($p<0.05$).

Clearly, the psychometric function when listening to L2 speech was generally shallower than when listening to L1 (Dutch) speech. For a second language for which the proficiency is lower (German compared to English), the mean of the distribution is not only shifted, but the steepness decreases as well. This is true at least for talkers E1M8 (English) and G1M5 (German); there is no reason to expect a different outcome for other talkers.

In terms of the 50% point of the psychometric function, nonauthentic pronunciation was found to be beneficial to Dutch listeners of German, but not of English (Fig. 1). Similar effects are not found on the slopes of the psychometric function.

IV. RELATIONS BETWEEN ACOUSTIC AND NONACOUSTIC FACTORS

A. The influence of context effects on SRT tests

In the case of non-native listeners, it seems likely that overall speech intelligibility is closely related to the listeners' skills at making use of linguistic redundancy (e.g., Bradlow and Pisoni; Bergman, 1980; Florentine, 1985; Mayo *et al.*, 1997). If this is true, we should be able to predict speech intelligibility from independent estimates of these linguistic skills. For this reason (if not for several others), it is worthwhile to look into methods of measuring listeners' use of linguistic redundancy.

A straightforward measure of linguistic redundancy is obtained through the letter guessing procedure (Shannon and Weaver, 1949), which uses orthographic presentations of sentences to obtain an estimate of linguistic entropy. Other suitable measures, such as the *j*- and *k*-factor by Boothroyd and Nittrouer (1988) and the *c*-parameters in the context model by Bronkhorst *et al.* (1993), require more complicated and cumbersome experiments.

B. Linguistic entropy (letter guessing procedure)

The letter guessing procedure (LGP) yields a measure of linguistic entropy (LE); this may be seen as the inverse of the effective redundancy through linguistic factors in the speech material. This measure has been used as a measure of individual subjects' linguistic skills (e.g., Van Rooij, 1991). Linguistic entropy has been shown to predict the influence of linguistic factors on speech intelligibility (Müsch and Buus, 2001; Van Rooij, 1991).

Since the procedure is based on orthographic presentations of test sentences, what it measures is by definition nonacoustic. Although it is possible to derive redundancy-related measures from spoken language tests, the LGP has some advantages. Because of the orthographic presentation, there are no individual talker effects, and the influence of speech acoustics is eliminated. Furthermore, redundancy at the subword level is included, since individual letters have to be guessed. For practical reasons, this is hard to achieve in any spoken language test, especially with non-native subjects. The orthographic approach also has clear disadvantages. Some factors that are irrelevant for spoken language intelligibility, such as spelling, are included. Also, some very relevant factors, such as phonological transition rules, are not incorporated in the test. However, it is fair to assume that linguistic entropy according to our definition may serve as an indicator of linguistic factors involved in speech recognition.

1. Subjects and stimuli

The subjects from groups I and II also participated in letter guessing procedure experiments. Although the same sentence material was used as in the SRT test, subjects were presented with each sentence in either the LGP or SRT test, but never saw or heard the same sentence more than once.

2. Procedure

The subject's task was to guess the next letter in an unfinished written sentence, displayed on a computer screen. The subject had to start out with no other information than an indication of the language of the next sentence, and had to guess the first letter using a computer keyboard.

After typing the guessed letter, the subject received visual and auditory feedback ("+" or "−" on the screen, high- or low-pitch sound). The correct letter was displayed on the screen, regardless of what the subject's response was. Next, the subject had to guess the next letter, following the same procedure (but with the added knowledge of what the first letter was). Letter by letter, the correct sentence appeared on the screen, while the subject responses, ignoring the difference between uppercase and lowercase, were stored.

The percentage of correctly guessed letters is a measure of linguistic redundancy. If a subject has no knowledge of the language whatsoever, he will guess each letter in a purely random fashion. Hence, in English he may statistically be expected to guess 1 out of 27 letters right (26 letters and

space). The more redundant the language is to the subject, the fewer letters he is forced to select randomly.

Rather than working directly with the percentage of correctly responded letters, the LGP scores are expressed in terms of linguistic entropy. Entropy, in the context of information theory, is expressed in ''bits.'' The linguistic entropy L is related to the fraction of correctly responded letters c according to[1]

$$L = -\log_2 (c). \tag{1}$$

Assuming a 27-letter alphabet (including space), the linguistic entropy associated with pure guessing of a single letter is, according to formula (1), 4.75 bits. This is the upper limit to L. If all letters are immediately guessed correctly, then $L = 0$: the material is perfectly redundant.

As an added measure, subjects were informally checked for their capacity to spell simple words in the tested language. For the letters that are particular to Dutch and German, not existing in English, the subjects were instructed to use similar characters that are usually assigned to replace these letters (e.g., ''ss'' for German ''β'').

Linguistic entropy will strongly depend on the type of sentences that are used: the more redundant the sentences, the smaller the estimated linguistic entropy. Even words within sentences will differ in terms of LE: semantic constraints will cause words towards the end of a sentence to be more redundant than words at the beginning of a sentence. When LE-estimates are calculated on a word-for-word basis, we expect the average LE as a function of the position of the word within sentences to be a monotonically decreasing function. For individual sentences this will usually not be true; in the phrase ''merry Christmas,'' for instance, the word ''Christmas'' is likely to be a local minimum in LE, regardless of the position within a sentence. However, when LE is measured as a function of word position across multiple sentences, differing somewhat in construction and number of words, a monotonically decreasing function seems likely. It also seems fair to assume that the LE decrease between two consecutive words becomes smaller toward the end of the sentence; the more context already exists, the smaller the gain will be by adding one extra word. When we assume that the LE decrease has an inverse proportional relation to word position n,

$$L_n - L_{n-1} = \frac{\alpha}{n}, \tag{2}$$

where $n \geqslant 2$ and α is an arbitrary constant, then L will be a function of n of the form

$$L_n = \beta + \alpha \ln n. \tag{3}$$

Here the constant β may be interpreted as the LE of a single word without sentence context; the constant α quantifies the effect of word position within a sentence on word LE. An exception is made for the first word ($n = 1$), for which Eq. (3) is not necessarily expected to hold. Within a set of sentences of a specific structure that is known to the subjects (such as SRT sentences), the predictability of the first word may be much higher than expected from Eq. (3).

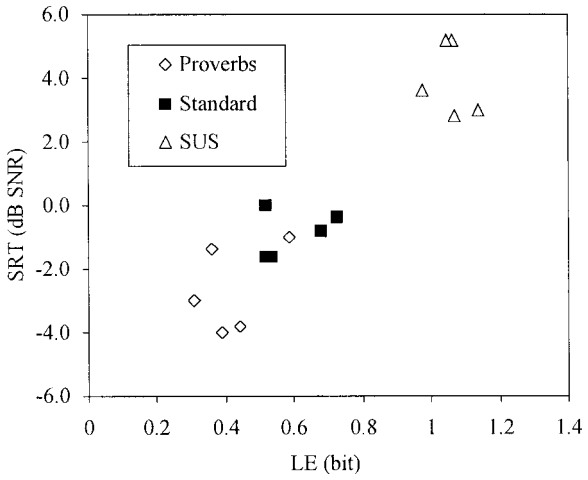

FIG. 5. Relation between SRT and LE, for five individual subjects and three types of SRT sentences. Results are mean values ($N = 2$ for SRT, $N = 13$ for LE). Speech material by the same talker was used for all SRT tests.

Since average LE effects due to word position will predominantly result from semantic constraints, semantic redundancy is in fact what the parameter α measures. By calculating LE as a function of word position across a sufficient number of subjects and sentences, the parameters α and β may be estimated using fixed nonlinear regression. By also estimating the standard errors associated with α and β, statistical significance is investigated by means of t-tests.

C. Results

1. Relation between LE and SRT for native speech communication

Linguistic entropy is the result of an interaction between subject and sentence material. If linguistic entropy estimates are to be used to quantify the effect of linguistic redundancy on SRT, this should also be possible in a fully native setting (Dutch subjects, Dutch language). The difference between subjects is then expected to be relatively small, but the amount of linguistic redundancy in the speech material can be varied systematically. This way, the relation between LE and SRT can be studied without introducing some of the uncertain factors that are automatically introduced when carrying out non-native perception experiments.

An important source of redundancy in natural speech is the use of semantic constraints. The SRT sentences form a homogeneous set in this respect. By constructing new sets of SRT sentences, which are designed to be as similar as possible to the ''standard'' SRT sentences in every way except semantic redundancy, the effect of semantic redundancy on native speech intelligibility may be evaluated. Similarly, the effect on linguistic entropy is investigated.

Two new sets of Dutch SRT sentences were constructed, one consisting of proverbs (higher than normal redundancy), the other consisting of semantically unpredictable sentences (lower than normal redundancy; Benoît *et al.*, 1996). LGP and SRT experiments were carried out with five native Dutch students, matching subject group II. Individual LE and SRT results are given in Fig. 5.

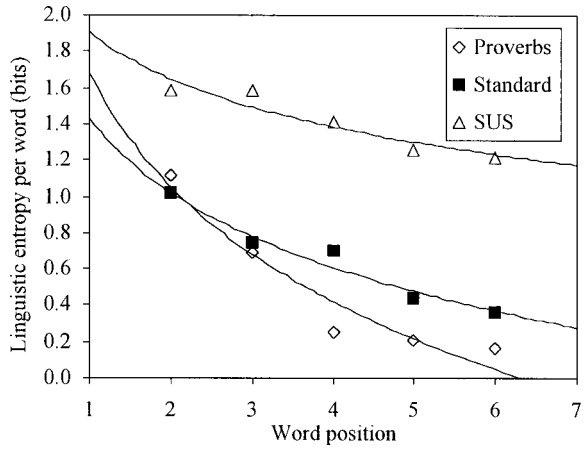

FIG. 6. Word-LE as a function of word position within sentences, for word positions $2 \leqslant n \leqslant 6$. The dashed lines are least-squares fits of Eq. (3) to the data for the three different kinds of sentences. Data points are based on five subjects (each 13 sentences) for proverbs and semantically unpredictable sentences, and on nine subjects (each 39 sentences) for the standard SRT sentences.

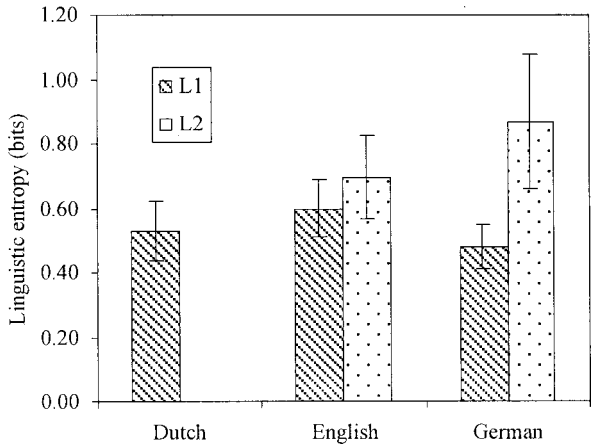

FIG. 7. Mean LGP results of L2 Dutch subjects (group I) and L1 German and American subjects. All L2 results and L1 Dutch results are based on nine listeners (39 sentences per listener, $N=351$); the L1 German and English results are based on three subjects (39 sentences, $N=117$). The error bars indicate the standard deviation.

Figure 5 shows some residual between-subject variance on the SRT scores, not explained by linguistic entropy. Still, the relation between SRT and LE across sentence types is clear. This means that differences in SRT can be predicted, to a certain degree, from linguistic entropy estimates. The mean increase in SRT as a function of LE is 10 dB/bit between the proverbs and the standard sentences. Between the standard sentences and the semantically unpredictable sentences, this slope is also 10 dB/bit.

The linguistic entropy of the three types of sentences was also calculated for individual words as a function of word position; results of this calculation are given in Fig. 6. The very first word of each sentence was not included in this analysis; its baseline predictability is much higher than all the other words, since it is nearly always an article.

Figure 6 shows that LE decreases monotonically with word position, as expected. The estimated values of parameters α and β from Eq. (3) are given in Table I.

If it is true that the three types of sentences differ primarily in semantic constraints, then we expect similar values of β, but different values for α. The differences in α are, as expected, statistically significant. However, the differences in β are also significant. This may indicate that, between the different sentence types, factors other than semantics were also different, such as word choice (mean frequency of occurrence in natural language, mean familiarity). It could also indicate that the assumption expressed by Eq. (2) is not completely justified for words at the beginning of sentences.

TABLE I. Estimated LE parameters from native LGP experiments for three types of sentences.

Sentence type	Slope (α)	Offset (β)	R^2 (explained variance)
Proverbs	−0.91	1.69	0.93
Standard SRT	−0.58	1.41	0.97
SUS	−0.38	1.91	0.88

2. Non-native LE results

With non-active listeners, linguistic entropy was not varied by manipulating the speech material; instead, it varied according to subjects' individual command of their second or third language. The LGP results of subject group I are presented in Fig. 7. Please note that the error bars in Fig. 7 indicate the standard deviation rather than the standard error, because of the large number of observations per language.

All differences in Fig. 7 are highly significant ($p < 0.001$). Unfortunately, and unlike the SRT results, the native (L1) LE scores are also significantly different between languages for L1 subjects. Hence, the LGP test is language dependent, and linguistic entropy estimates may not be compared across languages without applying corrections for differences in the LGP test.

The lowest native LE is found for German, then Dutch, and then English. The reduced entropy for German can be explained from a number of factors. Additional contextural constraints are introduced in German by the use of word gender and case, which is (virtually) not present in English, and of minor influence in Dutch. Moreover, the German convention of spelling nouns with capitalized first letters are also adopted in the feedback given by the LGP test, which also adds some redundancy.

Because of the differences between languages, we will use the ''normalized'' linguistic entropy from hereon. The normalization is accomplished by subtracting the mean native LE from the observed LE. This should largely eliminate between-language differences.

3. Relation between LE and SRT for non-native listeners

The effects of non-nativeness on LE appear to follow the same patterns as the SRT effects. This suggests that the overall intelligibility is largely determined by linguistic factors. Figure 8 shows the correlation between normalized LE and SRT for the individual subjects of group I+II (20 subjects) in all tested languages.

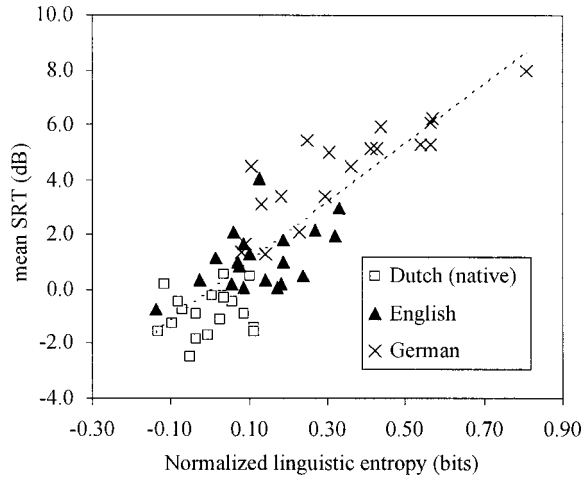

FIG. 8. Correlation between normalized LE and mean SRT (three talkers), for native Dutch and non-native English and German (20 subjects). All talkers were native in the given language. The dashed line is obtained through linear regression ($R^2=0.74$; slope 10.8 dB/bit, intercept −0.15 dB).

The value of the squared correlation coefficient ($R^2=0.74$) indicates that roughly 74% of the total variance in SRT scores in Fig. 8 may be explained using normalized linguistic entropy. This indicates that LE scores from letter guessing experiments can be used to obtain a fair prediction of corresponding SRT values.

More may perhaps still be learned from mean word LE as a function of word position, and by estimating the parameters α and β of Eq. (3). For the subjects of group I, we may verify the effect of the known difference in proficiency between (native) Dutch, English, and German (Fig. 9 and Table II).

All differences between the values of α and β in Table II are statistically significant. The influence of semantic constraints on LE, as quantified by slope α, is as could be expected for group I: apparently, the semantic constraints present in German sentences are not used as effectively as in English sentences.

The differences in β are not as easily interpreted, especially since β is higher for English than for German. If we assume that β expresses the linguistic entropy of words due to all factors other than semantic constraints, then this also includes the systematic differences between orthographic representations of the different languages. In this light, the fact that β is higher for English than for German does not seem as surprising anymore, but little room is left for interpretation of this parameter. Table II shows that group I subjects benefit more from semantic constraints in English than in German. However, although it appears likely that there is a relation with speech intelligibility, Table II does not provide information about this relation.

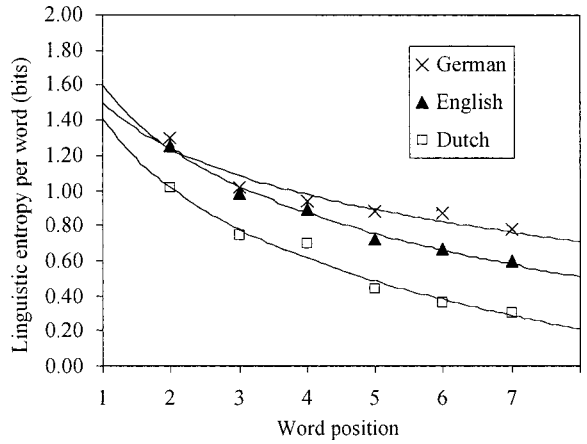

FIG. 9. Group I word-LE as a function of word position within sentences, for word positions $2 \leqslant n \leqslant 7$. The dashed lines are least-squares fits of Eq. (3) to the data for three different languages (native Dutch, and non-native English or German).

TABLE II. Estimated LE parameters from LGP experiments with group I subjects.

Sentence type	Slope (α)	Offset (β)	R_2 (explained variance)
Dutch (native)	−0.58	1.41	0.97
English	−0.52	1.60	0.99
German	−0.38	1.50	0.92

By investigating similar curves as given in Fig. 9 for groups of subjects differing in (non-native) speech intelligibility, the relation between the α parameter and the SRT may be established.

For the data presented in Fig. 10, the 20 subjects of group I+II were divided in four subgroups according to their mean SRT when listening to German by G1 talkers. For these subgroups of five subjects, word LE as a function of word position was calculated (Fig. 10 and Table III).

All differences between values of α and all differences between values of β are significant, with the exception of the differences for α and β for the 6.3- and 5.2-dB subgroups. This shows that intelligibility is related to the effective use of semantic constraints (α-parameter), as well as other linguistic factors (β-parameter).

V. DISCUSSION AND CONCLUSIONS

Using the speech reception threshold method, effects of non-native speech perception on speech intelligibility could

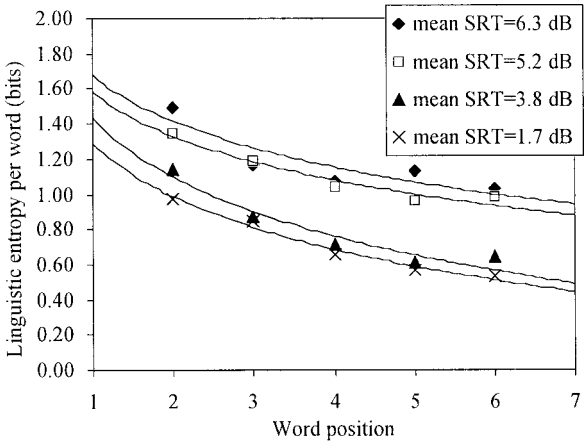

FIG. 10. Non-native German word-LE as a function of word position within sentences, for word positions $2 \leqslant n \leqslant 6$. The dashed lines are least-squares fits of Eq. (3) to the data for four subgroups of subject group I+II, differing in mean SRT (G1 speakers). Data points are based on five subjects (each 39 sentences).

TABLE III. Estimated LE parameters from LGP experiments with group I+II subjects (division into subgroups according to mean SRT scores for G1 talkers).

Mean SRT of subgroup (dB)	Slope (α)	Offset (β)	R_2 (explained variance)
6.3	−0.38	1.68	0.80
5.2	−0.36	1.58	0.94
3.8	−0.43	1.43	0.93
1.7	−0.48	1.29	0.98

be quantified for subjects ranging in proficiency from reasonable to excellent. Non-native speech recognition in noise does not just differ in terms of the mean of the psychometric function, but also the slope. To summarize the data given in this article, the average native (stylized) psychometric function and the worst-case non-native psychometric function derived from the experiments are given in Fig. 11.

The mean and slope of the psychometric functions of Fig. 11 can only be interpreted in the context of the specific sentence recognition paradigm used by the SRT test, implemented as described in this article. Other methods of measuring sentence recognition as a function of speech-to-noise ratio, or even other variations on the SRT paradigm, may lead to somewhat different results. For instance, relaxing the requirement that each individual word must be responded correctly will reduce the steepness of the curve. On the other hand, if optimized sets of selected test sentences are used (Versfeld *et al.*, 2000), then steeper psychometric functions will be found.

Despite the fact that there is a degree of dependency of the finding on the test method used, they also hold universal and quantitative meaning. If psychometric functions are known for two different test paradigms, in the same condition, then these curves can be used to transform measurement results from the scale of one test to the other. Hence, the difference between native and non-native intelligibility (given for our worst-case condition by the difference between both curves in Fig. 11) can also be transformed to other intelligibility scales, as long as the corresponding psychometric functions are known as a function of speech-to-noise ratio.

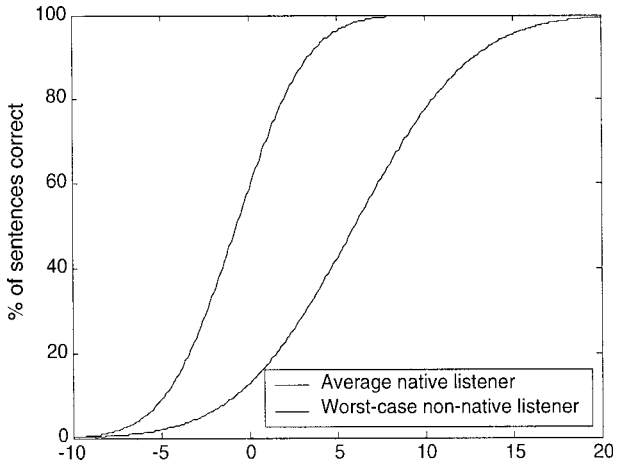

FIG. 11. Psychometric functions of speech reception in noise (percentage of sentences correctly received as a function of speech-to-noise ratio) for the average native listener from the SRT experiments (SRT=−0.7 dB, steepness 12.6%/dB) and the worst-case non-native listener (SRT=6.0, steepness 7.5%/dB).

A non-native listener with a degree of command on his second language that is better than that of the worst-case listener presented in Fig. 11 will produce a psychometric function when subjected to a SRT test that is somewhere between the two curves of Fig. 11.

For the listener populations and languages considered in this article, mean intelligibility effects of non-nativeness are sufficiently quantified by the outcome of the experiments. However, for other populations and languages, additional experiments will be needed. Carrying out listening experiments in non-native languages can be a time-consuming and difficult task. Letter guessing tests are easier to carry out, and the resulting linguistic entropy estimates predict speech intelligibility of non-native listeners with reasonable accuracy. This should open up possibilities to obtain (albeit somewhat crude) estimates of non-native listeners' intelligibility effects for a greater number of populations and languages.

As pointed out earlier in this work, the fact that linguistic entropy is a good predictor for intelligibility does not mean that the non-native speech recognition process is fully determined by linguistic factors. Since second-language learners tend to develop oral and written skills simultaneously, general second-language proficiency is an important explaining variable behind both linguistic entropy and SRT scores.

The fact that other than linguistic factors are also important is illustrated by the influence of L2 speech production (accented pronunciation) on L2 speech perception. Dutch listeners who were highly proficient in English experienced somewhat reduced speech intelligibility when listening to English by other non-native Dutch talkers, compared to native English talkers. For the same listeners, who were less proficient in German, the exact opposite was true for the German speech.

The experimental results offer no clear explanation for this discrepancy, but it seems that such an explanation is more likely to be found in the proficiency difference than in language-specific factors. The explanation could be that highly proficient listeners are able to use more subtle phonetic cues in authentically pronounced speech. The allophonic realizations of non-native talkers, even if they match the listeners' native model of phoneme space better, are less effective in transfering information needed in the speech recognition process. For less proficient listeners, these subtle phonetic cues are not as useful; they are unable to accurately categorize allophones using typically L2 phonetic contrasts, and perform better if these L2 allophones are ''mapped'' to their native phoneme space by non-native talkers.

In view of the results presented in Tables II and III, it seems likely that the contradictory findings by Florentine (1985) and others versus Koster (1987), regarding the use of semantic constraints by non-native listeners, can be explained by differences in their test population's mean proficiency. A high-proficiency population is likely to have ''near-

native'' use of contextual constraints, while this benefit is reduced for a low-proficiency population.

It is important to note that none of the experiments presented in this article were concerned with subjects of very poor proficiency. The earliest stages of second language learning may involve intelligibility effects beyond our scope of interest. However, people with sufficient command of a second language for practical daily usage will fall into categories somewhere between the two extremes given in Fig. 11. For the listener populations considered in this article, the presented measurement results can be used to assess exactly where between the lines in Fig. 11 we expect the psychometric function for a given population. For other languages and populations, additional data has to be collected. This data can consist of directly measured estimates of speech intelligibility; this is the best and most reliable option, but also the option that is the most difficult and time consuming. Alternatively, listeners' intelligibility effects can be predicted from measures that are easier to obtain, such as linguistic entropy estimates.

ACKNOWLEDGMENTS

The authors would like to thank Søren Buus and an anonymous reviewer for their very useful comments on an earlier version of this manuscript.

[1]Theoretically, linguistic entropy cannot be calculated from the fraction of correctly responded letters only, but needs to be corrected for the feedback (correct/incorrect) given to the subject. For simplicity, this correction is not included.

Benoît, C., Grice, M., and Hazan, V. (**1996**). ''The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences,'' Speech Commun. **18**, 381–392.

Bergman, M. (**1980**). *Aging and the Perception of Speech* (University Park, Baltimore, MD).

Boothroyd, A., and Nittrouer, S. (**1988**). ''Mathematical treatment of context effects in phoneme and word recognition,'' J. Acoust. Soc. Am. **84**, 101–104.

Bradlow, A. R., and Pisoni, D. B. (**1999**). ''Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors,'' J. Acoust. Soc. Am. **106**, 2074–2085.

Bronkhorst, A. W., Bosman, A. J., and Smoorenburg, G. F. (**1993**). ''A model for context effects in speech recognition,'' J. Acoust. Soc. Am. **93**, 499–509.

Buus, S., Florentine, M., Scharf, B., and Canevet, G. (**1986**). ''Native, French listeners' perception of American-English in noise,'' in *Proc. Internoise 86*, pp. 895–898.

Flege, J. E. (**1992**). ''The intelligibility of English vowels spoken by British and Dutch talkers,'' in *Intelligibility in Speech Disorders*, edited by R. D. Kent (Benjamins, Amsterdam).

Flege, J. E. (**1995**). ''Second-language speech learning: theory, findings, and problems,'' in *Speech Perception and Linguistic Experience*, edited by W. Strange (York, Baltimore, MD).

Flege, J. E., Bohn, O.-S., and Jang, S. (**1997**). ''Effects of experience on non-native speakers' production and perception of English vowels,'' J. Phonetics **25**, 437–470.

Florentine, M., Buus, S., Scharf, B., and Cavenet, G. (**1984**). ''Speech reception thresholds in noise for native and non-native listeners,'' J. Acoust. Soc. Am. Suppl. 1 **74**, S84.

Florentine, M. (**1985**). ''Non-native listeners' perception of American-English in noise,'' in *Proc. of Internoise 85*, pp. 1021–1024.

Gat, I. N., and Keith, R. W. (**1978**). ''An effect of linguistic experience; auditory word discrimination by native and non-native speakers of English,'' Audiology **17**, 339–345.

Koster, C. J. (**1987**). ''Word Recognition in Foreign and Native Language; Effects of Context and Assimilation,'' doctoral dissertation, University of Utrecht (Foris, Doredrecht, The Netherlands).

Lane, H. (**1963**). ''Foreign accent and speech distortion,'' J. Acoust. Soc. Am. **35**, 451–453.

Mayo, L. H., Florentine, M., and Buus, S. (**1997**). ''Age of second-language acquisition and perception of speech in noise,'' J. Speech Lang. Hear. Res. **40**, 686–693.

Meador, D., Flege, J. E., Mackay, I. R. A. (**2000**). ''Factors affecting the recognition of words in a second language,'' Bilingualism: Language and Cognition **3**, 55–67.

Müsch, H., and Buus, S. (**2001**). ''Using statistical decision theory to predict speech intelligibility. I. Model structure,'' J. Acoust. Soc. Am. **109**, 2896–2909.

Nábělek, A. K., and Donahue, A. M. (**1984**). ''Perception of consonants in reverberation by native and non-native listeners,'' J. Acoust. Soc. Am. **75**, 632–634.

Plomp, R., and Mimpen, A. M. (**1979**). ''Improving the reliability of testing the speech reception threshold for sentences,'' Audiology **18**, 43–52.

Shannon, C. E., and Weaver, W. (**1949**). *The Mathematical Theory of Communication* (Univ. of Illinois, Urbana).

Steeneken, H. J. M., and Houtgast, T. (**1999**). ''Mutual dependence of the octave-band weights in predicting speech intelligibility,'' Speech Commun. **28**, 109–123.

Strange, W. (**1995**). ''Cross-language studies of speech perception: A historical review,'' in *Speech Perception and Linguistic Experience*, edited by W. Strange (York, Baltimore, MD).

Van Rooij, J. C. G. M. (**1991**). ''Aging and the perception of speech: auditive and cognitive aspects,'' doctoral dissertation, Free University of Amsterdam.

van Wijngaarden, S. J. (**2001a**). ''The intelligibility of non-native Dutch speech,'' Speech Commun. **35**, 103–113.

van Wijngaarden, S. J., Steeneken, H. J. M., and Houtgast, T. (**2001**). ''Methods and models for quantitative assessment of speech intelligibility in cross-language communication,'' in *Proceedings of RTO Workshop on Multi-Lingual Speech and Language Processing*, Aalborg.

Versfeld, N. J., Daalder, J., Festen, J. M., and Houtgast, T. (**2000**). ''Method for the selection of sentence materials for efficient measurement of the speech reception threshold,'' J. Acoust. Soc. Am. **107**, 1671–1684.