Using the Speech Transmission Index for predicting non-native speech intelligibility

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

(Received 5 March 2003; revised 10 February 2003; accepted 15 December 2003)

While the Speech Transmission Index (STI) is widely applied for prediction of speech intelligibility in room acoustics and telecommunication engineering, it is unclear how to interpret STI values when non-native talkers or listeners are involved. Based on subjectively measured psychometric functions for sentence intelligibility in noise, for populations of native and non-native communicators, a correction function for the interpretation of the STI is derived. This function is applied to determine the appropriate STI ranges with qualification labels (''bad''–''excellent''), for specific populations of non-natives. The correction function is derived by relating the non-native psychometric function to the native psychometric function by a single parameter (ν). For listeners, the ν parameter is found to be highly correlated with linguistic entropy. It is shown that the proposed correction function is also valid for conditions featuring bandwidth limiting and reverberation.
 [DOI: 10.1121/1.1647145]

PACS numbers: 43.70.Kv, 43.71.Hw, 43.71.Gv [KWG] Pages: 1281–1291

I. INTRODUCTION

The intelligibility of speech is generally considered to depend on the characteristics of the talker and the listener, the complexity of the spoken messages, and the characteristics of the communication channel. Objective speech intelligibility predictions models have been shown to accurately predict the influence of the communication channel characteristics on speech intelligibility. An example of such a model is the Articulation Index (AI) model (French and Steinberg, 1947; Kryter, 1962), and more advanced models based on the AI, such as the Speech Intelligibility Index (SII; ANSI, 1997) and the Speech Transmission Index (STI; IEC, 1998; Steeneken and Houtgast, 1980; Steeneken and Houtgast, 1999).

In some cases, the overall speech intelligibility that is experienced is clearly affected by factors other than the physical characteristics of the channel. Individual talker differences (Bradlow *et al.* 1996; Hood and Poole, 1980) and message complexity (Pollack, 1964) were already mentioned. Other examples are individual differences in speaking style (Picheny *et al.* 1985) and hearing loss (Plomp, 1978).

An important determining factor for speech intelligibility is language proficiency, of talkers (van Wijngaarden *et al.*, 2002a) as well as listeners (van Wijngaarden *et al.*, 2002b). Learning a language at a later age results in a certain degree of limitation to language proficiency (Flege, 1995). So-called non-native speech communication is practically always less effective than native communication. The intelligibility effects of non-native speech production and non-native perception show an interaction with speech transmission quality (the quality of the channel). Speech degrading influences such as noise (Buus *et al.*, 1986; Florentine *et al.*, 1984; Florentine, 1985) and reverberation (Nábelek and Donahue, 1984) aggravate the intelligibility effects of non-native speech communication.

For various applications, it would be very useful to have an objective, quantitative intelligibility prediction method that is capable of dealing with non-native speech. In Sec. II of this article, the suitability of existing objective speech intelligibility prediction models for non-native applications is discussed.

Section III continues by proposing a way in which the Speech Transmission Index (STI) can be used in various non-native scenarios. Section IV contains a validation of this approach for speech in noise, bandwidth limiting, and reverberation.

II. SUITABILITY OF OBJECTIVE INTELLIGIBILITY PREDICTION MODELS FOR NON-NATIVE SPEECH

A. Speech transmission quality versus speech intelligibility

Speech intelligibility can be thought of as the success that a source and a receiver (talker and listener) have in transmitting information over a channel. Each unique talker–listener pair has a certain potential for transmitting messages of a given complexity. The quality of the transmission channel determines how much of this potential is realized. A typical transmission channel could be a phone line, a public address system, or the acoustic environment of a specific room.

Objective prediction models are especially good in quantifying speech transmission quality. The influence of factors determining speech intelligibility related to talkers and listeners, rather than the channel, has been incorporated to a lesser degree. A proficiency factor has been proposed (Pavlovic and Studebaker, 1984) for incorporating talker- and listener-specific factors into the framework of the articulation index, but this has not been developed to a level where practically useful predictions can be obtained.

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

TABLE I. Relation between STI and qualification labels.

Label	STI lower boundary	STI upper boundary
Bad	···	0.30
Poor	0.30	0.45
Fair	0.45	0.60
Good	0.60	0.75
Excellent	0.75	···

To predict the intelligibility of non-native speech, the interaction between speech transmission quality and language proficiency (quantified, for instance, by a linguistic entropy measure) of talkers and listeners needs to be studied.

B. Features of the SII, STI, and SRS models

At least three speech intelligibility prediction models presented in open literature show promise for predicting the effects of non-native factors: the Speech Intelligibility Index (SII; ANSI, 1997), the Speech Transmission Index (STI; IEC, 1998), and the Speech Recognition Sensitivity (SRS; Müsch and Buus, 2001a) models. Features of each separate model that are related to suitability for non-native applications are summarized in this section.

1. The Speech Transmission Index (STI)

The Speech Transmission Index combines the general concept of the articulation index with the observation that speech intelligibility is related to the preservation of the envelope spectrum of speech. The transmission quality of a channel is characterized by its modulation transfer function (MTF), which quantifies distortions in both the time and frequency domain (Houtgast *et al.*, 1980). The MTF is expressed as a matrix, giving a modulation index m as a function of 7 octave bands (125–8000 Hz) and 14 modulation frequencies (0.63–12.5 Hz). For conversational speech, using a wider range of modulation frequencies (up to 31.5 Hz) gives more accurate STI results in the presence of reverberation (van Wijngaarden and Houtgast, 2003).

The STI is purely a measure of speech transmission quality: it indicates to which degree the channel allows talkers and listeners to fulfill their potential for speech communication. Individual properties of talkers and listeners are not taken into account. The relation between STI and speech intelligibility has been verified and documented using various speech intelligibility measures (e.g., Houtgast and Steeneken, 1984).

To facilitate the use of the STI as an acceptability criterion, qualification labels (''bad''–''excellent'') have been attached to ranges of STI values (Table I). The ranges of Table I are based on the relation between STI and intelligibility for normal hearing, native subject populations, pragmatically taking ''round'' STI values as the category boundaries (ISO, 2002).

Commercially available measuring devices and measuring software can be used for *in situ* STI measures, or the STI can be calculated from theoretical knowledge of the channel (such as the output of room acoustics simulation software).

2. The Speech Intelligibility Index (SII)

The SII is an extension of a widely used version of the articulation index (Kryter, 1962) by incorporation the findings of Pavlovic, Studebaker, and others (e.g., Pavlovic, 1987; Pavlovic and Studebaker, 1984; Studebaker *et al.*, 1987). Instead of the MTF, the SII uses a band audibility function (based on the speech-to-noise ratio as a function of frequency) to quantify the contributions of different frequency bands to speech intelligibility.

The contribution of different frequencies to the SII is given by a frequency importance function. The ANSI standard associates different frequency importance functions with different measures of speech intelligibility. This means that the SII is not just a measure of speech transmission quality: it is designed to predict intelligibility according to different evaluation methods. Different SII values may be calculated for the same channel, depending on the chosen frequency importance function.

Poor communication is associated with SII below 0.45; good communication yields an SII in excess of 0.75.

3. The Speech Recognition Sensitivity (SRS) model

The SRS model, which uses statistical decision theory to explain how information is used across frequency, has quite recently been proposed, and has been shown to accurately predict intelligibility in a number of cases (Müsch and Buus, 2001a; 2001b). The SRS model explicitly includes listener-related factors that determine intelligibility, such as the power of ''cognitive noise'' that can be adjusted to fit the listener population. The predictability of the speech material (number of response alternatives in a recognition task) is also included in the model. The model can be applied to (qualitatively) explain the relation between linguistic entropy and speech intelligibility (see also Bronkhorst *et al.* 2002; van Rooij, 1991; van Wijngaarden *et al.*, 2002b). This is an attractive feature in the context of non-native speech communication, where linguistic entropy tends to be an important variable.

4. STI, SII, and SRS in relation to non-native speech

Of the prediction models described above, the SRS model is theoretically best equipped for dealing with non-native speech. Effects of non-native speech communication can be integrated directly through the model parameters. Despite the elegance of such a solution, the choice was made to base the approach proposed in this paper on the STI (in a manner to allow easy adaptation to the SII), not the SRS. The main reason is that, in order to make the results of our study as readily applicable as possible, a prediction method is sought that can be integrated seamlessly with tools already widely used to predict speech intelligibility, by researchers as well as engineers. The fact that the SRS method has (yet) to prove its validity and applicability as an operational tool outweighs, for the purposes of the current study, its theoretical appeal.

III. PROPOSED CORRECTION OF THE STI QUALIFICATION SCALE FOR NON-NATIVE SPEECH COMMUNICATION

A. Rationale for correcting the qualification scale

Modifying the STI method by including a proficiency factor (Pavlovic *et al.*, 1984) may seem attractive at first. It would change the index from a measure of speech transmission quality into more of an overall intelligibility predictor. However, the STI is commonly used to characterize communication channels (rooms or equipment), often for verification against certain minimum criteria (ISO, 2002). A talker-, listener-, or message-dependent STI may correlate better with intelligibility, but may also create confusion: the same channel can be characterized by various STI values, depending on factors other than the channel.[1]

We therefore propose to leave the STI calculation and measurement procedures unchanged. Instead, our approach is to make the *interpretation* of the STI dependent on language proficiency. This is done by correcting the qualification scale (Table I) for non-native speech communication. For each population of talkers and listeners, a specific correction applies, which makes sure that the qualification labels (''bad''–''excellent'') correspond to the same speech intelligibility as they normally do for native speech.

B. Method for correcting the qualification scale

1. Principles of the correction function

The key to relating the STI to non-native intelligibility lies in the difference between the psychometric functions for native and non-native speech recognition. The psychometric function $\pi(r)$ gives the percentage of correctly recognized test units (phonemes, words, or sentences), as a function of an independent variable r, which is a physical measure of speech degradation (such as speech-to-noise ratio, SNR). In cases where the independent parameter has a monotonic relationship with the STI, a correction function can be derived that relates a calculated or measured (''native'') STI, to a ''non-native STI'' that is required to obtain the same intelligibility in case of non-native communication. This correction function can then be applied to the qualification scale boundaries, relating the standard STI to the proper qualification labels for non-native communication. Please note that the correction function is used to calculate the *required* STI to achieve a certain level of intelligibility, not to change the STI value itself.

Figure 1 is a visual representation of a correction function, where the independent variable r is the speech-to-noise ratio. The noise spectrum is presumed to be equal to the long-term average speech spectrum, and no speech degrading influences other than noise are present. This results in a simple relation between STI and SNR, represented by the double horizontal axis labeling. The L1 and L2 psychometric curves in Fig. 1 are fictitious. Intelligibility qualifications (Table I) represent different levels of intelligibility (the vertical axis in Fig. 1). By following the arrows, the required native STI to reach a certain level of intelligibility is translated into a required non-native STI, that corresponds to the same intelligibility.

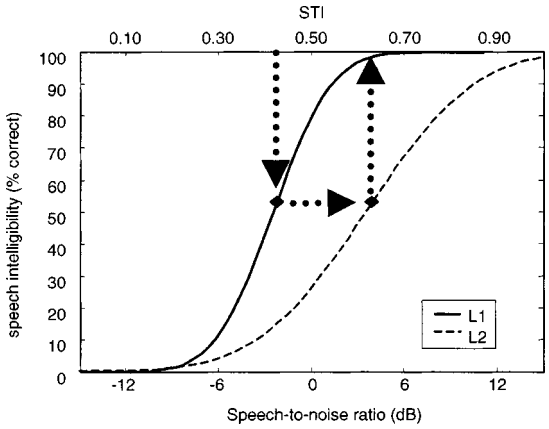

FIG. 1. Schematic representation of the procedure for deriving a correction function for non-native interpretation of the STI. The psychometric curves are fictitious, but representative of those found when measuring native and non-native sentence intelligibility.

Functions $f(r)$ to calculate the STI for different choices of physical parameter r, such as bandwidth, speech-to-noise ratio (SNR), and reverberation times, are known. The operations visualized by Fig. 1 can only be carried out mathematically if the relation $f(r)$ is reversible, meaning that Eq. (1) must be a unique function

$$r = f^{-1}(\mathrm{STI}). \tag{1}$$

This is, for instance, the case for additive noise that has the same long-term spectrum as speech, provided that no other speech degrading factors are present (the case of Fig. 1). The SNR then fully determines the STI, so each value of the STI corresponds to a single SNR.[2] All that is needed to calculate a correction function is a model of the psychometric functions shown in Fig. 1. Of the possible choices for independent variable r, the SNR is the easiest and most directly accessible option, and will be used throughout this paper.

After mathematically deriving (or numerically implementing) the correction of Fig. 1, it can be applied to the STI boundaries of Table I. For each population of L2 talkers and listeners, the correction function will be different, leading to specific versions of Table I.

2. Deriving the correction function from psychometric function models

Assuming that the psychometric function for native (L1) speech may be approximated by a cumulative normal distribution (e.g., Versfeld *et al.*, 2000), it is best described by

$$\pi_{L1}(r) = \Phi\left(\frac{r-\mu_{L1}}{\sigma_{L1}}\right), \tag{2}$$

where $\Phi(z)$ is the standardized cumulative normal distribution, μ_{L1} and σ_{L1} are the mean and standard deviation of the distribution for fully native speech, respectively. A straightforward way to derive a correction function is to assume that Eq. (2) also holds for non-native speech, in which case μ_{L2} and σ_{L2} will depend on the average proficiency level of the population. By solving $\pi_{L1} = \pi_{L2}$, substituting Eq. (1), a correction function as given in Eq. (3) is obtained

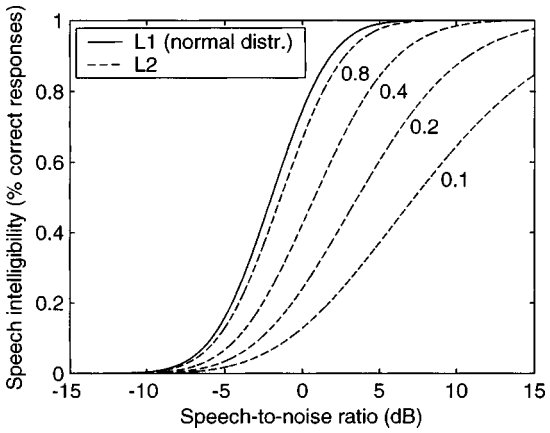

FIG. 2. Examples of L2 psychometric functions derived from a cumulative normal L1 psychometric function ($\mu=-2,\sigma=3$), according to Eq. (4), for $\nu=0.8$, $\nu=0.4$, $\nu=0.2$, and $\nu=0.1$.

$$\mathrm{STI}_{L2}=f\left(\sigma_{L2}\left(\frac{f^{-1}(\mathrm{STI}_{L1})-\mu_{L1}}{\sigma_{L1}}\right)+\mu_{L2}\right). \tag{3}$$

Thus, assuming that, for a certain type of test that measures intelligibility as a function of r, μ_{L1}, and σ_{L1} are known, the information needed to correct a required STI_{L1} into an equivalent required STI_{L2} is a specification of the L2 population in terms of μ_{L2} and σ_{L2}.

Earlier results show that μ_{L2} and σ_{L2}, when estimated as two separate parameters, are not independent. They tend to be highly correlated: when the mean of the psychometric function shifts, the slope also changes. This is related to the behavior of L1 and L2 psychometric functions near 0% intelligibility. In all cases, intelligibility starts to ''build up'' from 0% around the same SNR, for listeners (van Wijngaarden *et al.*, 2002b: Fig. 11) as well as talkers (van Wijngaarden *et al.*, 2002a: Fig. 6). In other words, L1 and L2 psychometric curves share a common origin (in Fig. 1 around -12 dB). The most likely reason is that the detection threshold for L1 and L2 speech is the same; hence, contributions to intelligibility are expected from the same SNR (the detection threshold) upward. However, as the SNR increases, intelligibility rises more quickly for L1 than L2 subjects, causing the psychometric functions to diverge. This suggests that, instead of estimating the two parameters μ_{L2} and σ_{L2}, the L2 psychometric function can be derived from the L1 psychometric function using a single parameter ν, according to Eq. (4)

$$\pi_{L2}(r)=1-(1-\pi_{L1}(r))^{\nu}. \tag{4}$$

The parameter ν (cf. Boothroyd and Nittrouer, 1988) can assume any value between 0 (no speech recognition at all) and 1 (native speech communication), and quantifies the degree to which non-native intelligibility is able to keep up with native intelligibility as the SNR increases, from the detection threshold upward. A family of psychometric functions according to Eq. (4), derived from a L1 psychometric function that follows a normal distribution, is shown in Fig. 2.

It appears that Eq. (4) describes earlier experimental data very well, with only one parameter (ν) instead of two (μ_{L2} and σ_{L2}), while allowing a very intuitive interpretation. Another advantage has to do with artifacts at low SNRs when calculating the STI correction function. Small errors in estimates of μ and σ may lead to an L2 psychometric function that is locally higher than the L1 function. Although the difference in intelligibility at these SNRs is very small, the effect on the correction function according to Eq. (3) can be noticeable.

A disadvantage of Eq. (4) is that a correction equation cannot be obtained in mathematically closed form by simply solving $\pi_{L1}=\pi_{L2}$, if the L1 psychometric curve is modeled as a cumulative normal distribution [Eq. (1)]. Sometimes the logistic function is used as an approximation of the cumulative normal distribution (e.g., Versfeld *et al.*, 2000). In that case, the correction function in closed form can be calculated (see the Appendix). However, due to differences around the tails of the distribution, small but noticeable deviations in the calculated correction function are observed compared to a correction function based on the cumulative normal distribution.

A numerical implementation of the correction function as a function of ν was easily realized, based on Eqs. (1), (2), and (4), following the procedure visualized in Fig. 1. This numerical implementation was used to calculate the correction functions used in this study.

3. Complexity of test material to use for measuring psychometric functions

Message complexity and context effects are always key factors for speech intelligibility (Pollack, 1964), but especially when non-native listeners are involved. Context effects influence speech intelligibility differently for non-natives than for natives (e.g., Mayo *et al.*, 1997; van Wijngaarden *et al.*, 2002b). This means that a correction function as visualized in Fig. 1 depends on the amount of contextual information in the test material.

Our aim for the correction function is to allow interpretations of the STI for non-natives in the same way as for natives, in practical situations where non-native talkers or listeners are involved. This means that the test material used to obtain correction functions must contain the same sources of contextual information that are also expected in practice (telephone conversations, public address messages, etc.). Correction functions based on, for instance, psychometric curves for phoneme recognition would have little practical meaning; differences in use of contextual information would simply not be included in the correction. A suitable choice of test material, representative of common situations involving non-natives, seems to be a corpus of everyday sentences, carrying a representative amount of semantic, syntactic, and lexical redundancy.

The corrections used in this paper are all based on psychometric functions obtained using an implementation of the Speech Reception Threshold (SRT) procedure (Plomp and Mimpen, 1979). The SRT is the SNR at which the intelligibility of short, redundant sentences is 50%. Additional measurements, at fixed SNRs around the SRT, were used to estimate the slope of the psychometric function (van

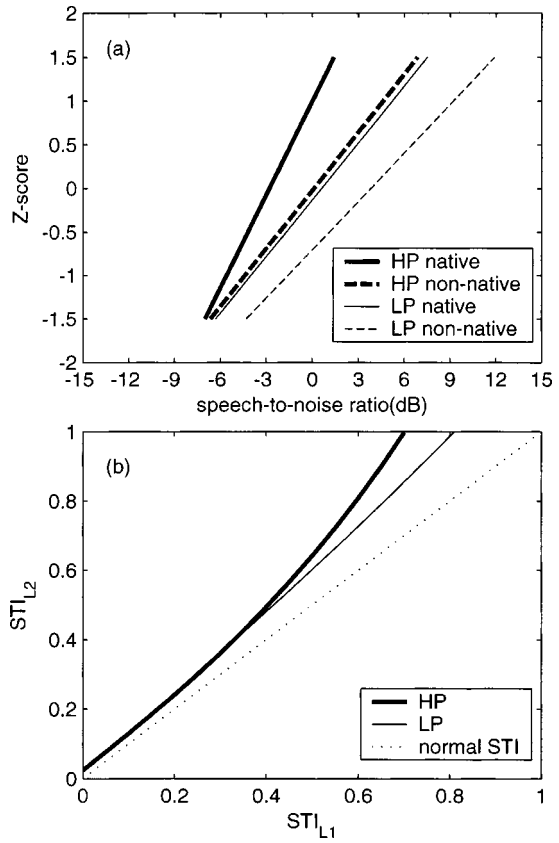

FIG. 3. (a) Psychometric functions, in terms of Z-score as a function of SNR, for high-predictability (HP; $\nu=0.36$, $\mu_{L1}=-2.8$ dB, and $\sigma_{L1}=2.8$ dB) and low-predictability (LP; $\nu=0.50$, $\mu_{L1}=0.6$ dB, and $\sigma_{L1}=4.6$ dB) sentences (after Florentine, 1985); (b) the STI correction functions derived from these psychometric functions.

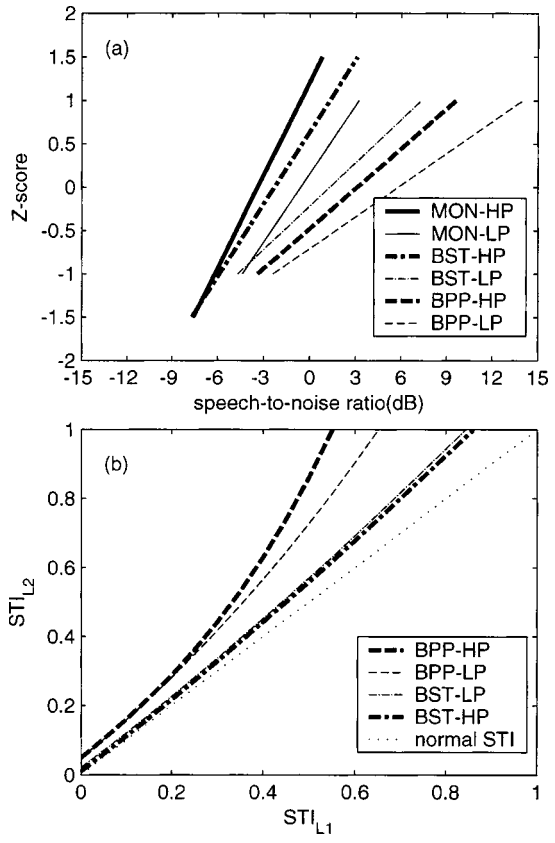

FIG. 4. (a) Psychometric functions, in terms of Z-score as a function of SNR, for high-predictability (HP; $\mu_{L1}=-3.4$ dB and $\sigma_{L1}=2.8$ dB) and low-predictability (LP; $\mu_{L1}=-0.5$ dB and $\sigma_{L1}=3.8$ dB) sentences, for three groups of nine subjects: monolinguals (MON), early bilinguals (bilingual since toddler, BST), and late bilinguals (bilingual post puberty, BPP; after Mayo *et al.*, 1997); (b) the STI correction functions derived from these psychometric functions.

Wijngaarden *et al.*, 2001). The speech recordings that were used were part of the VU corpus (male talker) of SRT sentences (Versfeld *et al.*, 2000).

C. Qualification labels for non-native listeners

1. Correction functions for different populations of listeners

To summarize the previous section: a correction of the qualification scale can be derived from any study that results in native and non-native intelligibility of everyday sentences, as a function of SNR. Several studies yielding such results for non-native listeners have been reported.

Florentine (1985) used the Speech Perception in Noise (SPIN) test (Kalikow and Stevens, 1977) to measure intelligibility of high-predictability (HP) and low-predictability (LP) sentences, with a mixed population of 16 non-native subjects. Results were compared to similar results for 13 native (U.S. English) listeners. The final word in HP sentences was semantically predictable, the final word in LP sentences was not. Scoring was based only on recognition of the final word. This makes the HP sentences a more suitable candidate for deriving a correction function; since semantic redundancy is important for practical non-native scenarios, it should be reflected by the correction function.

The original data taken from Florentine (1985) are shown in Fig. 3(a). From the reported psychometric functions (given as Z-scores as a function of SNR), separate values of μ_{L1} and σ_{L1} were taken for HP and LP sentences, and values of ν were obtained using a Gauss–Newton nonlinear fitting procedure. The correction functions for HP ($\nu=0.36$) and LP ($\nu=0.50$) sentences are given in Fig. 3(b).

The difference between correction functions for high-predictability and low-predictability sentences is clear. The difference in ν can be seen as a quantification of Florentine's finding that non-natives are not as able as natives to make use of semantic redundancy.

Following an approach similar to Florentine's, Mayo *et al.* (1997) investigated speech perception of Mexican–Spanish-speaking listeners in English. Groups of early bilinguals (bilingual-since-toddler, BST) and late bilinguals (bilingual-post-puberty, BPP) were compared to native English subjects using the SPIN test.[3] All groups consisted of nine subjects. The original data are given in Fig. 4(a), the derived correction functions in Fig. 4(b).

The correction functions differ between early bilinguals ($\nu=0.64$ for HP, $\nu=0.57$ for LP) and late bilinguals ($\nu=0.15$ for HP, $\nu=0.22$ for LP). The proficiency differences are reflected by differences in ν, and in relation to that, by the slope of the correction function.

Earlier data from trilingual non-native listeners (van Wijngaarden *et al.*, 2002b) yield similar results for ν values

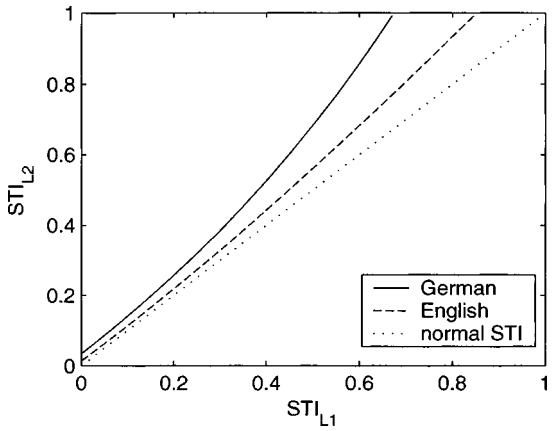

FIG. 5. STI correction functions for trilingual Dutch listeners of German (low proficiency, $\nu=0.21$) and English (high proficiency, $\nu=0.52$); $\mu_{L1}=-0.50$ dB and $\sigma_{L1}=3.21$ dB (after van Wijngaarden *et al.*, 2002b).

and correction functions as the data by Mayo *et al.* The trilingual subjects were highly proficient in English, and showed poor to moderate proficiency in German. The SRT sentence material used to obtain these results is closest to the HP sentences of the SPIN test. Calculated mean ν values are 0.21 (German speech) and 0.52 (English speech). The corresponding STI correction functions are given in Fig. 5.

2. Relation between STI and qualification labels for non-native listeners

By applying the correction functions of Figs. 3, 4, and 5 to Table I, the STI qualification label boundaries of Table II are obtained. From Figs. 3 and 4, the functions for HP sentences are used.

Table II shows how qualitative descriptions of populations of listeners, such as early versus late bilinguals, or low-proficiency versus high-proficiency listeners, can be used for the interpretation of the STI. The same speech transmission quality (STI) leads to different qualifications of intelligibility, depending on the population of listeners.

The SRT data behind Fig. 5 can also be related to L2 listeners' proficiency in a quantitative way (van Wijngaarden *et al.*, 2002b). Along with SRT results, estimates of linguistic entropy were obtained using the letter guessing procedure (LGP; Shannon and Weaver, 1949; van Rooij, 1991). This orthographic procedure, which measures the extent to which subjects are able to make use of linguistic redundancy, can be seen as a measure of proficiency, which correlates well with non-native speech intelligibility. A strong relation between linguistic entropy and the ν parameter is expected. Linguistic entropy and psychometric function estimates were obtained separately, using different subject groups (which where matched for L2 proficiency, age, and gender). Unfortunately, this means that LGP results from that study cannot be related to the ν parameter on an individual level. However, the mean linguistic entropy L can be compared to the mean value of ν for three different languages: native Dutch ($L=0.53$, $\nu=1$ by definition), English ($L=0.70$, $\nu=0.57$), and German ($L=0.87$, $\nu=0.23$). The explained variance by correlating these data ($R^2=0.995$), if only on the basis of three observations, seems promising.

To further investigate this relation, new experiments were carried out with eight native and eight non-native listeners. The non-native group consisted of L2 learners of the Dutch language, with different language backgrounds (American English, Amharic, German, Greek, Hungarian, Indonesian, Polish, and Tigrinya) and different levels of proficiency. All were late bilinguals, differing mainly in L2 experience. Six of the listeners could be classified as relatively low-proficiency subjects, with average of 4 years of experience with the Dutch language, and a mean self-reported proficiency (on a five-point scale) of 3.2. The other two subjects were classified as high proficiency, with an average of 13 years of experience, and a self-reported proficiency of 4.5. The native group was matched to the non-native group in terms of age, gender, and level of education. All subjects were between 19 and 33 years of age, and were taking part in (or had recently completed) higher education in the Netherlands.

In order to be able to estimate the ν parameter for the non-native subjects, individual psychometric functions were measured for all 16 listeners. Sentences in noise were presented at five fixed SNRs, centered around the SRT with 2-dB intervals. The mean percentage of correctly recognized sentences was measured using 13 sentences per SNR, after which the psychometric function was fitted. This procedure was repeated three times with each listener; the mean of these three fits was taken to obtain a more accurate estimate.

For the native subjects, the psychometric function was assumed to be a cumulative normal distribution. The mean native psychometric function in this experiment is described by $\mu_{L1}=-4.38$ dB and $\sigma_{L1}=2.20$ dB. For each individual non-native listener, the psychometric function was related to the mean native psychometric function according to Eq. (4), by fitting the ν parameter.

A significant correlation was found between linguistic

TABLE II. Relation between STI and qualification labels for non-native listeners, after correction according to Figs. 4 and 6 (HP sentences), and Fig. 7. The text ">1" indicates that an STI greater than 1 would be required, meaning that this qualification cannot be reached.

STI label category boundary	Standard	Florentine (1985)	Mayo *et al.* (1997) BST (early)	Mayo *et al.* (1997) BPP (late)	van Wijngaarden *et al.* (2002) English	van Wijngaarden *et al.* (2002) German
Bad–poor	0.30	0.36	0.33	0.44	0.33	0.38
Poor–fair	0.45	0.57	0.50	0.74	0.50	0.60
Fair–good	0.60	0.79	0.68	>1	0.68	0.86
Good–excellent	0.75	>1	0.86	>1	0.87	>1

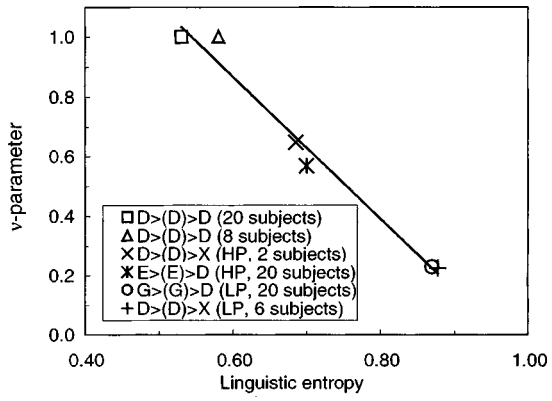

FIG. 6. Relation between mean linguistic entropy and the ν parameter, for Dutch learners of German and English (20 subjects) and learners of Dutch from various language backgrounds (two high-proficiency listeners, six low-proficiency listeners; $R^2=0.98$).

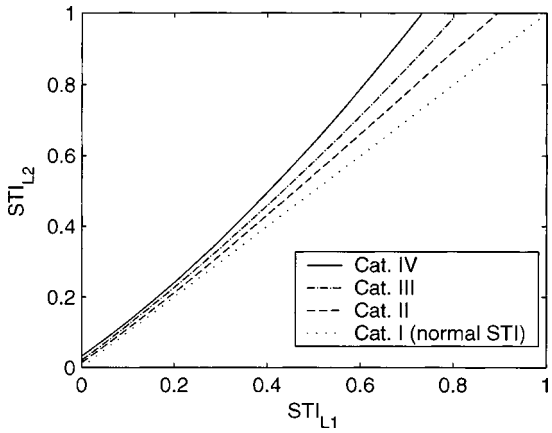

FIG. 7. STI correction functions for L2 talkers of the Dutch language, for different degrees of foreign accent strength (cat. I–cat. IV; van Wijngaarden *et al.*, 2002a). Category I means that the talker has (virtually) no foreign accent, category IV means that the accent is severe (see Table III for the corresponding values of ν, μ_{L1}, and σ_{L1}).

entropy and the ν parameter on an individual level ($R^2=0.74$). The means of the native, high-proficiency, and low-proficiency subjects in this experiment are given in Fig. 6, along with the means from the earlier experiments in German (low proficiency) and English (high proficiency).

As seen in Fig. 6, linguistic entropy estimates are found in the 0.50–0.60 range for native subjects; linguistic entropy is higher for non-natives. Despite the differences in test languages and language backgrounds of the listeners, the data from the two experiments seem to fit the same relation between linguistic entropy and the ν parameter. The importance of this relation lies in the fact that the experimental procedures to determine a subject's linguistic entropy requires only a fraction of the time needed to assess the ν parameter on an individual basis. Through the ν parameter, the interpretation of the STI for non-natives can be derived from linguistic entropy estimates.

D. Qualification labels for non-native talkers

Psychometric functions describing the intelligibility of foreign-accented speech are similar to the ones observed for non-native listeners, although non-native speech production tends to have a smaller overall impact on speech intelligibility than non-native perception.[4] Previously reported data on talkers from four different categories of accent strength (numbered I–IV, ranging from "native" to "severe accent") were used to calculate STI correction functions (Fig. 7). The resulting STI label categories are given in Table III.

Figure 7 and Table III are based on data obtained with native listeners. Translation of the STI to objective qualification labels when non-native talkers *and* non-native listeners are involved is not possible using Tables II and III, due to interaction effects. The overall intelligibility may be higher than expected when simply adding up the individual effects of non-native talking and non-native listening. This so-called interlanguage speech intelligibility benefit may occur when the native language of talker and listener is the same, but also if their language background is different (Bent and Bradlow, 2003).

IV. VALIDATION OF THE QUALIFICATION SCALE CORRECTION

A. Validation issues

If speech is degraded by additive, steady-state noise only, there is little reason to question the validity of the correction functions described above. With the already-mentioned limitations regarding the amount of contextual information in the intelligibility test material, the approach of correcting the required STI for a certain level of intelligibility (by finding the STI value that leads to equal intelligibility for non-native communication) should work by definition. However, in the presence of speech degrading influences other than steady-state noise, the validity of this approach remains to be proven. Two important sources of speech degradation are bandwidth limiting and reverberation.

Using the STI correction functions for non-native speech communication in cases where the SNR depends on frequency implies the assumption that the relative importance of all frequency bands is the same as for native speech. The validity of this assumption is verified by measuring speech intelligibility of bandwidth-limited speech in noise for non-native and native listeners.

In case of reverberation, the STI model expressed the degree of speech degradation in terms of an "equivalent speech-to-noise ratio," which is calculated through the modulation transfer function (MTF). Again, the correction function approach is only valid under the assumption that

TABLE III. Relation between STI and qualification labels for non-native talkers differing in degree of foreign accent, after correction according to Fig. 7. The text ">1" indicates that an STI greater than 1 would be required, meaning that this qualification cannot be reached. The mean ν value for each category is also given ($\mu_{L1}=-0.50$ dB, $\sigma_{L1}=3.21$ dB).

STI label boundary	Category: Standard STI (Cat. I)	Cat. II (ν=0.67)	Cat. III (ν=0.48)	Cat. IV (ν=0.32)
Bad–poor	0.30	0.32	0.34	0.36
Poor–fair	0.45	0.49	0.52	0.56
Fair–good	0.60	0.66	0.71	0.79
Good–excellent	0.75	0.85	0.91	>1

this MTF-based operation is equally valid for non-native as for native communicators. To investigate this, speech intelligibility is measured under reverberant conditions, with native and non-native listeners.

Once intelligibility measurements in bandwidth-limited and reverberant conditions have been carried out, there is a straightforward procedure to investigate whether the validity of the proposed correction functions extends to these conditions. The correction functions are based on measures of speech intelligibility as a function of STI (Fig. 1). However, the only independent parameter [r in Eq. (1)] that was varied to obtain different values of the STI was the speech-to-noise ratio. When bandwidth limiting and reverberation come into play, the relation between intelligibility and STI (native and non-native) must remain the same as the noise-only case for the correction functions to remain valid.

In other words: regardless of the type of degradation, a certain level of intelligibility (such as 50% intelligibility of sentences) must always correspond to the same STI. This was one of the design objectives for the STI method, and normally found to be true for native speech (Steeneken and Houtgast, 1980). For the proposed correction functions to be valid, the same must be true for non-native speech. Maintaining the same, consistent relation between (corrected) STI and speech intelligibility (the same for bandwidth limiting and reverberation as for noise-only) is a necessary and sufficient condition for validity.

B. Effects of bandwidth limiting

The same 16 listeners who participated in the SRT and LGP experiments reported above and shown in Fig. 6 took part in an experiment consisting of SRT measurements in bandwidth-limited conditions. The experiments were carried out in Dutch, using the eight Dutch subjects to obtain a native baseline. The eight non-native listeners were treated as a single group, and were all presented with the same conditions as the native listeners. SRT sentences pronounced by a single male Dutch speaker were used, in a wideband condition as well as three bandwidth-limited conditions. The bandwidth-limited conditions offered a bandwidth of 4 octaves (500-Hz–4-kHz bands), 3 octaves (500-Hz–2-kHz bands) and 2 octaves (1-kHz and 2-kHz bands). Complementary stop-band noise was added to the bandlimited speech, to prevent spreading of information into adjacent bands through nonlinear auditory phenomena.

In each of the conditions, the SRT was measured (the SNR corresponding to 50% sentence intelligibility). The corresponding STI was calculated, based on the available bandwidth and the SNR resulting from the SRT measurement. Because the SRT is the SNR corresponding to a fixed level of intelligibility (namely 50%), the ''STI at the SRT'' should be a constant value for the proposed correction function approach to be valid; as indicated in the previous section, this is a necessary and sufficient condition for validity. Results are given in Fig. 8.

For non-native as well as native listeners, the STI at the SRT is fairly constant. With the exception of the difference between the wideband and the three-octave condition for the native group, none of the within-group differences in Fig. 8 is statistically significant ($p<0.05$). The average across all bandwidth-limited conditions (native and non-native considered separately) does not differ significantly from the wideband condition. This means that the proposed approach is also valid for bandwidth-limited conditions. To further illustrate this, Table IV compares corrected native STI values to the non-native STI values as shown in Fig. 8. The close correspondence between corrected native STI and measured non-native STI demonstrates that the correction function can indeed be used in conditions featuring noise as well as bandwidth limiting.

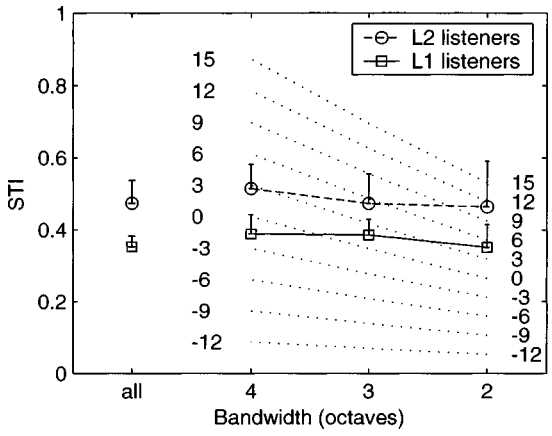

FIG. 8. STI at the SRT, for conditions with and without bandwidth limiting. The dotted lines indicate the maximum STI at each bandwidth, as a function of the SNR. The errorbars indicate the standard deviation ($N=24$; 8 listeners, each 3 SRT measurements per condition).

The mean native STI results fall in the range between 0.30 and 0.45, leading to a classification of ''poor'' according to the standard table (Table I). The mean non-native results for each condition would be (incorrectly) categorized as ''fair.''

The ν value for each non-native listener was determined in a separate experiment, following the procedure described above in relation to Fig. 6. Using the mean value of the ν parameter across all L2 listeners ($\nu=0.33$), a correction function for this population of non-native listeners was obtained. After applying this correction function, the L2 results correctly fall into the ''poor'' category (the corresponding STI range after correction is $0.37<\mathrm{STI}<0.59$).

C. Effects of reverberation

In addition to bandwidth-limiting conditions, SRT experiments were carried out in conditions featuring reverberation. The same subjects participated, and speech material by the same talker was used.

TABLE IV. ''STI at the SRT'' results with and without bandwidth limiting, for non-native and native subjects. STI means and s.d.'s are calculated across 8 subjects, 3 observations per condition.

	L1 STI		L1 STI after correction		L2 STI	
Condition	Mean	s.d.	Mean	s.d.	Mean	s.d.
Wideband	0.35	0.03	0.44	0.05	0.47	0.07
4 octaves	0.39	0.05	0.50	0.07	0.51	0.07
3 octaves	0.39	0.04	0.50	0.06	0.47	0.08
2 octaves	0.35	0.06	0.44	0.09	0.46	0.13

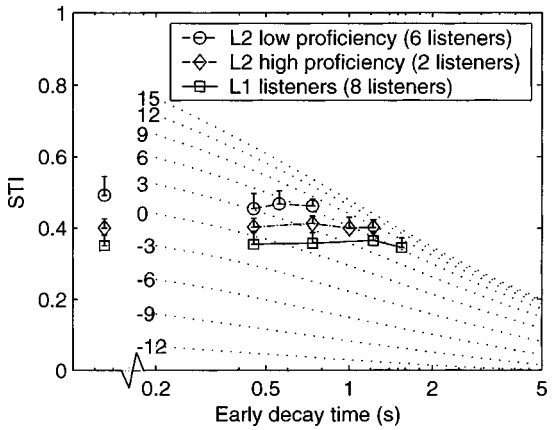

FIG. 9. STI at the SRT, for conditions with and without reverberation. The dotted lines indicate the maximum STI at each EDT, as a function of the SNR. The errorbars indicate the standard deviation (2 to 8 listeners, each 3 SRT measurements per condition). The EDT in this plot is the mean EDT in the octave bands 125 Hz–8 kHz. The STI calculation is nonstandard, and includes modulation frequencies up to 31.5 Hz.

To obtain conditions differing in early decay time (EDT), but with as little other differences as possible, the same highly reverberant room was used for all conditions. The only difference between conditions was the amount of acoustic absorption material in the room. Impulse responses with a length of approx. 1.5 s were recorded in each condition, and stored digitally. From these impulse responses, the EDT was measured in each octave band.

In order to be able to present reverberant speech to the subjects without physically having to change the acoustic properties of the reverberant room between conditions, the prerecorded impulse responses were used for the stimulus presentations. The SRT test sentences were convolved with the impulse responses in real time, using an overlap-add procedure. All stimuli were presented diotically, excluding binaural effects (for which the STI has not been validated) from the experiment. For the experiment, conditions with EDTs between approx. 0.5 and 2 s were used. The eight native subjects all participated in the same conditions. The differences in proficiency between the L2 subjects were such that some were able to carry out the test at longer EDTs than others. For this reason, the same distinction between ''high proficiency'' (two subjects) and ''low proficiency'' (six subjects) used in Fig. 6, was again applied. Results of STI calculations at the SRT as a function of EDT, similar to Fig. 8, are given in Fig. 9.

The STI calculations underlying Fig. 9 are based on a modulation frequency range of 0.63–31.5 Hz instead of the standardized range (0.63–12.5 Hz), for reasons related to speaking style and envelope spectrum of the talker (van Wijngaarden and Houtgast, 2003). In earlier, similar experiments concerned with the effects of reverberation, the ''STI at the SRT'' was found to be independent of early decay time for normal hearing as well as hearing impaired listeners (Duquesnoy and Plomp, 1980).

For all three groups in Fig. 9, the STI at the SRT appears to be independent of EDT, and (nearly) the same as for the condition without reverberation. The mean values for the reverberant conditions do not differ significantly from the condition without reverberation. This indicates that the same STI always represents the same level of intelligibility, in noise as well as reverberation, meaning that the proposed correction function approach is valid for reverberant conditions as well. This is illustrated by Table V, which shows that only a small difference remains between the measured non-native STI and the native STI after correction.

V. DISCUSSION AND CONCLUSIONS

A. The ν parameter

The approach for non-native interpretation of the STI, as proposed in this article, is based on a few novel concepts. Perhaps the most important of these is modeling the non-native psychometric function by relating it to the native psychometric function, through a single parameter ν. This has several advantages, such as its intuitive interpretation, and the fact that this parameter can be related to linguistic entropy (which can be measured with relative ease). Among the disadvantages of this approach is the fact that the non-native psychometric function, even when derived from a native function that *is* modeled as a cumulative normal distribution, does not exactly follow such a normal distribution itself. This causes mathematical complications, and may take away some of its theoretical appeal. However, measurements of the non-native psychometric function appear to be in support of this psychometric function model. The particular way in which differences in proficiency result in a family of psycho-

TABLE V. ''STI at the SRT'' results with and without reverberation, for non-native and native subjects. STI means and s.d.'s are calculated across 8 subjects, 3 observations per condition. The early decay time given in this table is the mean EDT across octave bands 125 Hz–8 kHz.

Early decay time (s)	L1 STI		High proficiency: L1 STI after correction		High proficiency: L2 STI		Low proficiency: L1 STI after correction		Low proficiency: L2 STI	
	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
(no rev.)	0.35	0.03	0.38	0.03	0.40	0.03	0.48	0.05	0.49	0.05
0.59	0.35	0.04	0.39	0.04	0.40	0.03	0.49	0.05	0.45	0.05
0.75	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	0.47	0.04
1.00	0.36	0.03	0.39	0.03	0.41	0.02	0.49	0.04	0.46	0.02
1.22	⋯	⋯	⋯	⋯	0.40	0.02	⋯	⋯	0.44	0.02
1.76	0.36	0.03	0.40	0.04	0.40	0.02	0.51	0.04	⋯	⋯
2.62	0.35	0.03	0.37	0.03	⋯	⋯	0.47	0.03	⋯	⋯

metric curves (e.g., van Wijngaarden *et al.*, 2002a: Fig. 6) closely matches expectations based on differences in the ν parameter. This leads us to conclude that this non-native psychometric function model is the most appropriate choice for our current purposes.

B. Effects of linguistic message content

Our correction function approach yields, by definition, representative results if the only speech degrading factor is steady-state noise, and if the messages have the approximate linguistic characteristics of SRT sentences. This indicates two specific concerns for the validity of the approach: differences in complexity of the speech material, and speech degrading conditions other than additive noise. Section IV dealt with the concerns regarding other types of speech degradation. Message complexity is an issue that perhaps needs closer consideration; differences were found between correction functions for high-predictability (HP) and low-predictability (LP) sentences, indicating that differences in semantic redundancy can result in different correction functions (Figs. 3 and 4). However, the STI is most commonly applied to situations where little variation in semantic redundancy is expected. Moreover, deviations between the HP and LP curves only appear to occur for subjects of quite low proficiency, and then only on the high end of the STI scale. In conclusion, if reasonably representative sentence material is chosen for measurement of the psychometric curves, than the specific details of linguistic content are considered to be of minor importance. In Figs. 3 and 4, the HP curves are expected to be most representative of the STI application domain.

C. Application of the proposed approach

Any prediction of speech intelligibility for a population of non-native talkers or listeners must always be based on some description of this population. Preferably, this should be a description in terms of easily observed or accessible characteristics (such as a general categorization of L2 proficiency, or severity of foreign accent). The approach outlined in this article is based on the use of systematically measured psychometric functions, matched with some of these observations and characteristics (specifically accent ratings and linguistic entropy).

As an efficient procedure for obtaining a correction function for non-native listeners, one could estimate the linguistic entropy distribution for the target population using the letter guessing procedure (Shannon and Weaver, 1949). This is a time-efficient procedure; it is feasible to collect distributions of individual linguistic entropy for larger populations of non-native listeners, for instance, by setting up a booth at an international airport, or even through the Internet. Once a distribution of linguistic entropy for the target population is known, the next step is an external choice: how do we wish to represent this population? The mean of the distribution will be appropriate for many applications, while for some, one may want to choose a more conservative threshold (for instance, the 25th percentile, in which case 75% of the population shows equal or better proficiency than the threshold). Using the relation shown in Fig. 6, the chosen entropy threshold can be converted into the equivalent value of the ν parameter, from which the corresponding correction function can be calculated.

For talkers, a similar approach can be adopted, but based on a distribution of proficiency self-ratings rather than linguistic entropy. Combined with a categorization scheme such as the one used in Fig. 7, self-ratings can also be translated into equivalent values of the ν parameter.

In conclusion, the proposed correction function approach broadens the scope of applicability of the STI method to include various applications involving non-natives. Obvious applications include public address systems at airports, and auditoria used for international conferences.

APPENDIX: DERIVATION OF AN STI CORRECTION FUNCTION BASED ON A LOGISTIC FUNCTION

Deriving a correction function based on the psychometric functions described by Eqs. (2) and (4) involves solving $\pi_{L1}=\pi_{L2}$, as represented by Eq. (A1)

$$\Phi\left(\frac{r_{L1}-\mu_{L1}}{\sigma_{L1}}\right)=1-\left[1-\Phi\left(\frac{r_{L2}-\mu_{L1}}{\sigma_{L1}}\right)\right]^{\nu}. \tag{A1}$$

The cumulative normal distribution $\Phi([r-\mu]/\sigma)$ may be approximated by a logistic function (e.g., Versfeld *et al.*, 2000), such as Eq. (A5)

$$\Lambda(\rho)=\frac{e^{\rho}}{1+e^{\rho}}, \tag{A2}$$

where

$$\rho=\frac{r-\mu}{\sigma\sqrt{\pi/8}}. \tag{A3}$$

By substituting $\Lambda(\rho)$ for $\Phi([r-\mu]/\sigma)$ in Eq. (A1) and solving, Eq. (A4) is obtained

$$\rho_{L2}=\ln[(e^{\rho_{L1}}+1)^{(1/\nu)}-1]. \tag{A4}$$

By substituting Eqs. (1) and (A3) in Eq. (4), the correction function Eq. (A5) is obtained

$$\mathrm{STI}_{L2}=f(\mu_{L1}+\sigma_{L1}\sqrt{\pi/8}\ln[(e^{[f^{-1}(\mathrm{STI}_{L1})-\mu_{L1}]/(\sigma_{L1}\sqrt{\pi/8})}+1)^{(1/\nu)}-1]). \tag{A5}$$

[1]Something similar *always* applies (even without using the proficiency factor) for the SII, since the SII depends on the type of intelligibility test it aims to predict.

[2]This is only true if the SNR is between −15 and +15 dB. Outside this range, the STI is (respectively) always 0 or 1, meaning that STI=1 corresponds to any SNR greater or equal than +15. This topic is addressed later on in this section.

[3]Mayo *et al.* (1997) also tested a separate bilingual-since-infancy (BSI) group. Because of the limited number of subjects in this group (3), Mayo *et al.* chose to combine their BST and BSI groups for statistical analysis. The BSI group data are not used in this article.

[4]This statement is based on comparisons of SRT results between cases where only the talker is non-native, and only the listener is non-native (talkers and listeners of comparable proficiency). In both cases, the speech material is fixed; this means that the non-native talkers do not rely on their own linguistic resources (vocabulary, syntactical knowledge, etc.), but simply use the language that is handed to them. If the dynamics of free conversation are taken into consideration, the situation will be much more

complex, and the comparison between the magnitudes of perception and production effects may have a different outcome.

ANSI (**1997**). ANSI S3.5-1997, ''Methods for calculation of the speech intelligibility index'' (American National Standards Institute, New York).

Bent, T., and Bradlow, A. R. (**2003**). ''The interlanguage speech intelligibility benefit,'' J. Acoust. Soc. Am. **114**, 1600–1610.

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., Toretta, G. M., and Pisoni, D. B. (**1996**). ''Intelligibility of normal speech. I. Global and fine-grained acoustic-phonetic talker characteristics,'' Speech Commun. **20**, 255–272.

Bronkhorst, A. W., Brand, T., and Wagener, K. (**2002**). ''Evaluation of context effects in sentence recognition,'' J. Acoust. Soc. Am. **111**(6), 2874–2896.

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.

Duquesnoy, A. J. H. M., and Plomp, R. (**1980**). ''Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis,'' J. Acoust. Soc. Am. **68**, 537–544.

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

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

Florentine, M., Buus, S., Scharf, B., and Canevet, G. (**1984**). ''Speech reception thresholds in noise for native and non-native listeners,'' J. Acoust. Soc. Am. **75**, S84–S84.

French, N. R., and Steinberg, J. C. (**1947**). ''Factors governing the intelligibility of speech sounds,'' J. Acoust. Soc. Am. **19**, 90–119.

Hood, J. D., and Poole, J. P. (**1980**). ''Influence of the speaker and other factors influencing speech intelligibility,'' Audiology **19**, 434–455.

Houtgast, T., and Steeneken, H. J. M. (**1984**). ''A multi-lingual evaluation of the Rastimethod for estimating speech intelligibility in auditoria,'' Acustica **54**, 185–199.

Houtgast, T., Steeneken, H. J. M., and Plomp, R. (**1980**). ''Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics,'' Acustica **46**, 60–72.

IEC (**1998**). IEC 60268-16 2nd ed. ''Sound system equipment. Part 16: Objective rating of speech intelligibility by speech transmission index'' (International Electrotechnical Commission, Geneva, Switzerland).

ISO (**2002**). ISO/FDIS 9921 ''Ergonomics—Assessment of speech communication'' (International Organization for Standardization, Geneva, Switzerland).

Kalikow, D. N., and Stevens, K. N. (**1977**). ''Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability,'' J. Acoust. Soc. Am. **61**, 1337–1351.

Kryter, K. D. (**1962**). ''Methods for the calculation and use of the articulation index,'' J. Acoust. Soc. Am. **34**, 1689–1697.

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.

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

Müsch, H., and Buus, S. (**2001b**). ''Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance,'' J. Acoust. Soc. Am. **109**(6), 2910–2920.

Nábelek, 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.

Pavlovic, C. V. (**1987**). ''Derivation of primary parameters and procedures for use in speech intelligibility predictions,'' J. Acoust. Soc. Am. **82**, 413–422.

Pavlovic, C. V., and Studebaker, G. A. (**1984**). ''An evaluation of some assumptions underlying the articulation index,'' J. Acoust. Soc. Am. **75**, 1606–1612.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1985**). ''Speaking clearly for the hard of hearing I. Intelligibility differences between clear and conversational speech,'' J. Speech Hear. Res. **28**, 96–103.

Plomp, R. (**1978**). ''Auditory handicap of hearing impairment and the limited benefit of hearing aids,'' J. Acoust. Soc. Am. **63**(2), 533–549.

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

Pollack, I. (**1964**). ''Message probability and message reception,'' J. Acoust. Soc. Am. **36**, 937–945.

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

Steeneken, H. J. M., and Houtgast, T. (**1980**). ''A physical method for measuring speech transmission quality,'' J. Acoust. Soc. Am. **67**, 318–326.

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

Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (**1987**). ''A frequency importance function for continuous discourse,'' J. Acoust. Soc. Am. **81**, 1130–1138.

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

van Wijngaarden, S. J., and Houtgast, T. (**2004**). ''Effect of talker and speaking style on the Speech Transmission Index,'' J. Acoust. Soc. Am. **115**(1), 38–41.

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 the RTO Workshop on Multi-lingual Speech and Language Processing* (Aalborg, Denmark).

van Wijngaarden, S. J., Steeneken, H. J. M., and Houtgast, T. (**2002a**). ''Quantifying the intelligibility of speech in noise for non-native talkers,'' J. Acoust. Soc. Am. **112**(6), 3004–3013.

van Wijngaarden, S. J., Steeneken, H. J. M., and Houtgast, T. (**2002b**). ''Quantifying the intelligibility of speech in noise for non-native listeners,'' J. Acoust. Soc. Am. **111**(4), 1906–1916.

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.