# Quantifying the intelligibility of speech in noise for non-native talkers

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 28 February 2002; accepted for publication 8 August 2002)

The intelligibility of speech pronounced by non-native talkers is generally lower than speech pronounced by native talkers, especially under adverse conditions, such as high levels of background noise. The effect of foreign accent on speech intelligibility was investigated quantitatively through a series of experiments involving voices of 15 talkers, differing in language background, age of second-language (L2) acquisition and experience with the target language (Dutch). Overall speech intelligibility of L2 talkers in noise is predicted with a reasonable accuracy from accent ratings by native listeners, as well as from the self-ratings for proficiency of L2 talkers. For non-native speech, unlike native speech, the intelligibility of short messages (sentences) cannot be fully predicted by phoneme-based intelligibility tests. Although incorrect recognition of specific phonemes certainly occurs as a result of foreign accent, the effect of reduced phoneme recognition on the intelligibility of sentences may range from severe to virtually absent, depending on (for instance) the speech-to-noise ratio. Objective acoustic-phonetic analyses of accented speech were also carried out, but satisfactory overall predictions of speech intelligibility could not be obtained with relatively simple acoustic-phonetic measures.  [DOI: 10.1121/1.1512289]

PACS numbers: 43.70.Kv, 43.71.Hw, 43.71.Gv [KRK]

## I. INTRODUCTION

The intelligibility of a speech utterance depends on many factors, among which the individual characteristics of the talker. Differences between the intelligibility of individual talkers are caused by, among other things, differences in articulatory precision (Bradlow *et al.*, 1996), speaking rate (Sommers *et al.*, 1994), and speaking style (Picheny *et al.*, 1985; Bradlow and Pisoni, 1999). A special class of talker characteristics stems from being raised in another language than the language that is being spoken. These characteristics cause listeners to perceive the speech as foreign accented; moreover, they may reduce the intelligibility of the speech.

The effect of non-nativeness on speech intelligibility sometimes complicates communication with non-native talkers significantly. Especially under adverse conditions, such as background noise and bandwidth limiting, non-native talkers tend to be less intelligible (e.g. Lane, 1963; van Wijngaarden, 2001a).

Knowing the extent to which the intelligibility of non-native talkers is reduced can be very useful. Predictions of speech intelligibility are widely used in systems design and engineering; for instance, for the design of telecommunication equipment and in room acoustics. When the influence of having a non-native talker on speech intelligibility can be quantified, design criteria can be adjusted.

Of course, having a foreign accent will not affect speech intelligibility equally for all non-native talkers. Experienced second language talkers, and talkers who started learning their second language at a relatively early age, are likely to suffer a smaller decrease in speech intelligibility (e.g., Flege *et al.*, 1997). By conducting speech intelligibility experiments for closely defined populations of talkers (in terms of all relevant factors, including L2 experience and age of acquisition) it should be possible to quantify intelligibility effects of non-nativeness for these populations. Preferably, one would like to be able to predict speech intelligibility effects from talker characteristics that are easily observed.

In order to properly quantify speech intelligibility effects, it is essential that out of many ''standard'' methods to measure intelligibility, a method is chosen that is suitable for quantifying effects of non-nativeness (van Wijngaarden, 2001b). In principle, segmental as well as supra-segmental influences can be expected. There has traditionally been much attention to effects found at the phoneme level. Researchers find more or less consistent patterns of phoneme confusions, largely depending on the relation between the language background of talkers and listeners (e.g., Peterson and Barney, 1952; Singh, 1966). Although the occurrence of these confusions will surely reduce the overall intelligibility, it is unclear *to what degree*. The presence of context will enable listeners to correctly interpret many nonauthentic speech sounds, despite the talker's poor production.

It seems reasonable to expect that the overall effect of non-nativeness on speech intelligibility is closely related to the degree of perceived foreign accent. Not unlike the degree of perceived accent, the overall effect on speech intelligibility results from several characteristics of non-native speech production. Without examining all of these characteristics in detail, one would expect that the degree of foreign accent would predict the effect on speech intelligibility, and vice versa. This hypothesis can be tested by examining speech intelligibility and foreign accent for talkers, differing in L2 proficiency.

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

 

TABLE I. Measures related to the foreign accent of 15 speakers of the Dutch language. The mean proficiency self-rating is the mean across four different self-ratings (speaking, listening, reading, and writing). The pairwise comparison rating is derived from an experiment in which 19 native listeners compared all combinations of the 15 talkers presented in this table, in a total of 39 sessions.

| Talker | Native language | Age of first acquisition | Experience with Dutch (years) | Self-rating for speaking | Mean self-rating | Pairwise comparison rating (overall foreign accent) |
|---|---|---|---|---|---|---|
| DM-1 | Dutch | ⋯ | ⋯ | 5 | 5 | −1.80 |
| DM-2 | Dutch | ⋯ | ⋯ | 5 | 5 | −1.61 |
| DF-3 | Dutch | ⋯ | ⋯ | 5 | 5 | −1.50 |
| GM-4 | German | 23 | 3 | 4 | 4.25 | −0.05 |
| GM-5 | German | 28 | 0.5 | 2 | 3 | 1.01 |
| GF-6 | German | 19 | 11 | 4 | 4 | −1.07 |
| EF-7 | Am. English | 23 | 6 | 3 | 3.25 | 0.02 |
| EM-8 | Am. English | 19 | 28 | 5 | 4.75 | −0.78 |
| EM-9 | Am. English | 27 | 2.5 | 2 | 3.25 | 0.99 |
| PM-10 | Polish | 24 | 2 | 3 | 2.5 | 0.65 |
| PF-11 | Polish | 26 | 2 | 2 | 2.5 | 1.36 |
| PF-12 | Polish | 26 | 1.5 | 2 | 2.5 | 0.72 |
| CF-13 | Chinese | 20 | 21 | 4 | 3.5 | −0.59 |
| CF-14 | Chinese | 23 | 0.25 | 2 | 2 | 1.22 |
| CF-15 | Chinese | 27 | 20 | 2 | 2 | 1.44 |

The objective of this study is to find a way to quantify the effects of a non-native talker on speech intelligibility. The relative importance of low-level (phoneme) and high-level (sentence) effects of non-native speech production on intelligibility is examined. Furthermore, the relationship between accent and speech intelligibility is investigated, hoping to establish a method to predict speech intelligibility from accent strength. The reliability of non-native talkers' self-ratings for their second language proficiency is also determined.

Under perfect listening conditions, even subjects with a strong accent can be perfectly intelligible. As communication conditions become more adverse (due to speech degrading factors such as additive noise, bandwidth limiting, or reverberation) the effects of foreign accent on speech intelligibility can be expected to increase. For this reason, the experiments in this study are all concerned with speech in the presence of *noise*. The influence of noise can be seen as representative for many speech degrading conditions.

## II. DEGREE OF PERCEIVED FOREIGN ACCENT

### A. Methods

Inexperienced second language (L2) talkers are often recognized as being non-native because their L2 speech production incorporates typical traits of their native language. The resulting foreign accent is usually perceived holistically, despite the fact that certain specific deviations from native speech production can be pointed out (e.g., Magen, 1998; Flege, 1984). The components that constitute a foreign accent are both segmental (such as deviations from expected voice onset times, effects of poorly developed L2 phonetic categories) and supra-segmental (less authentic intonation, unnatural pauses, effects on speaking rate). Upon being presented with non-native speech fragments of sufficient length, native listeners should be able to produce foreign accent ratings that include influences of all relevant cues.

One could reason that non-native talkers can hardly be reliable judges of their own accent. The reasons why non-native talkers exhibit a certain accent are certain limitations of their L2 speech production. These limitations may perhaps also be expected to affect (or even originate from) speech perception, rendering them ''deaf'' to certain aspects of their own accent.

However, this does not mean that non-native talkers' self-ratings for their second language proficiency are useless. Our main interest in the degree of foreign accent comes from the hypothesis that this may predict the extent to which speech intelligibility is affected. Proficiency self-ratings by non-native talkers may serve the same purpose, even if these talkers are not sensitive to their own accent. It seems reasonable to assume that non-native talkers are aware of their own proficiency in producing second-language speech, because of the fact that they are repeatedly confronted with the effects of their accent. Especially non-native talkers that are submerged in an L2 environment should be able to assess the strength of their own accent, if only by its apparent effect on native listeners.

#### *1. Subjects, method for obtaining self-ratings*

Speech recordings were made for a total of 15 talkers. Three of the talkers were native Dutch, the other 12 were learners of the Dutch language from four different language backgrounds (German, English, Polish, and Chinese; three talkers for each language background). The talkers also differed with respect to gender, age of acquisition, time since the first contact and average frequency of use of the Dutch language (Table I).

All talkers were asked to rate their Dutch proficiency on a five-point scale, assigning separate ratings for their oral and written skills, both passive (reading/listening) and active (speaking/writing).

All self-ratings were registered just before the start of a speech recording session. The talkers were given the oppor-

tunity to revise their self-ratings after the recording session, but none of the talkers chose to do so.

### *2. Method for obtaining accent ratings from pairwise comparisons*

In order to obtain accurate accent ratings with a relatively limited number of native listeners, a pairwise comparison experiment was carried out. The listeners compared each voice out of the set of 15 talkers to every other voice, always indicating which of the two showed the strongest foreign accent. Computer-stored speech samples of at least 15 s in length were presented to the listeners through headphones, by means of a high-quality sound device. The listeners were allowed to repeat speech samples of the pair of talkers as often as they liked, switching back and forth between the voices as they wished. They could indicate which of the two had the strongest accent by pressing buttons on a computer keyboard.

Upon completion of the experiment by a listener, a preference matrix was compiled from the results. By adding such matrices across multiple subjects, an average preference matrix (representing the preferences of the listener group as a whole) was composed. To extract accent ratings from the preference matrix, this matrix was converted to a probability matrix and subjected to a *Z*-transform. By then adding all elements in each column (or row) of the matrix a rating of the subjective accent strength was obtained (Torgerson, 1958).

The sentences used in the experiment were taken from the speech reception threshold (SRT) corpus (Plomp and Mimpen, 1979), and recorded using the procedure designed for creating a multi-lingual SRT database (van Wijngaarden, 2001b). The same sentences were used for both voices in each pair.

A total of 19 native listeners participated; ten of these listeners repeated the experiment three times with different speech material. Hence, all ratings are based on 39 sets of comparisons between all talkers. All listeners were between 17 and 31 years of age, and tested for having normal hearing.

## B. Results

In Table I, relevant information regarding the 15 talkers is given, together with proficiency self-ratings and accent ratings from the pairwise comparison experiment.

As can be seen in Table I, the L2 talkers differ with respect to their experience with the Dutch language. All first started learning Dutch as an adult. Hence, the experimental results obtained with these talkers apply to clearly postlingual second language learners.

One would expect a decrease of the degree of foreign accent with L2 experience. Such a relationship is already informally observed in Table I, and further established by Fig. 1, which shows the foreign accent rating by native listeners as a function of the number of years of experience with the Dutch language. Talker CF-15 takes an exceptional position. This talker reported 20 years of L2 experience, but was also the only talker to indicate a very low frequency of use of the Dutch language; she was also the only talker without written Dutch skills.

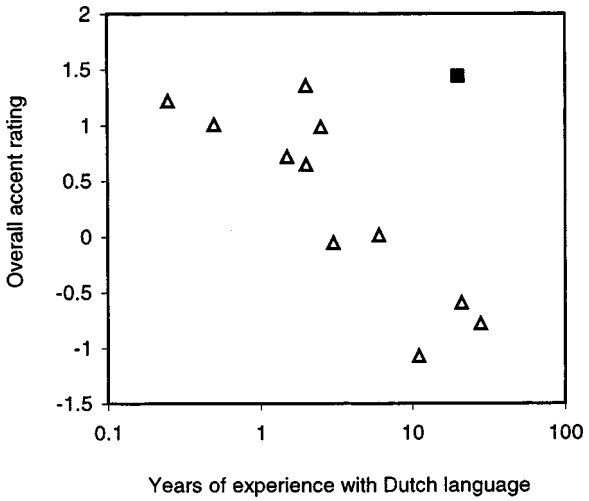


FIG. 1. Relation between foreign accent ratings and years of experience with the Dutch language, for the 12 L2 talkers. With the exception of talker CF-15 (indicated by a black square) the accent rating correlates well with the logarithm of the number of years of experience ($R^2 = 0.74$, without CF-15).

Please note the logarithmic scale in Fig. 1. The degree of foreign accent decreases with experience, but this decrease slows down as a function of time.

To investigate the correlation between self-ratings for speaking proficiency and foreign accent rating by native listeners, these measures are plotted against each other in Fig. 2.

The correlation between self-ratings and foreign accent is relatively strong; 91% of the total variance in foreign accent strength can be accounted for from self-ratings only.

We are mostly interested in the degree of foreign accent for its effect on speech intelligibility. In this light, a limitation of the accent ratings from Figs. 1 and 2 is that, since the subjects rated accent holistically, various speech characteristics may have attributed to the ratings. For example, a fluent

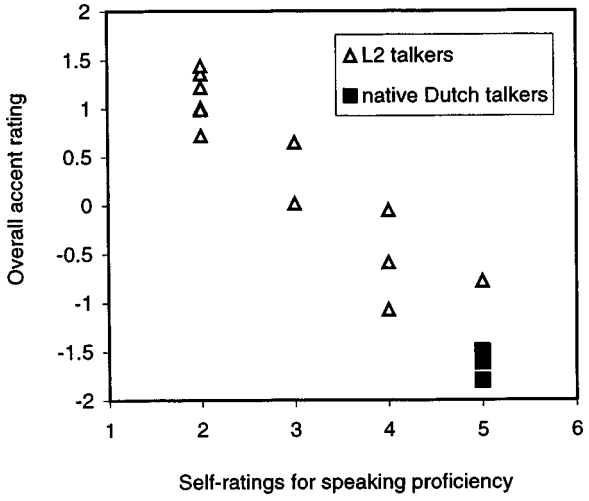


FIG. 2. Relation between self-ratings for speaking proficiency and foreign accent ratings from pairwise comparisons by native Dutch listeners ($R^2 = 0.91$).

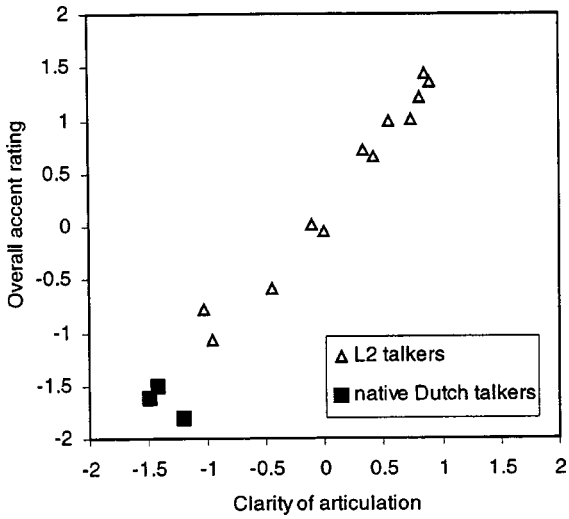


FIG. 3. Relation between pairwise comparison ratings for "clarity of articulation" and overall foreign accent ($R^2=0.97$).

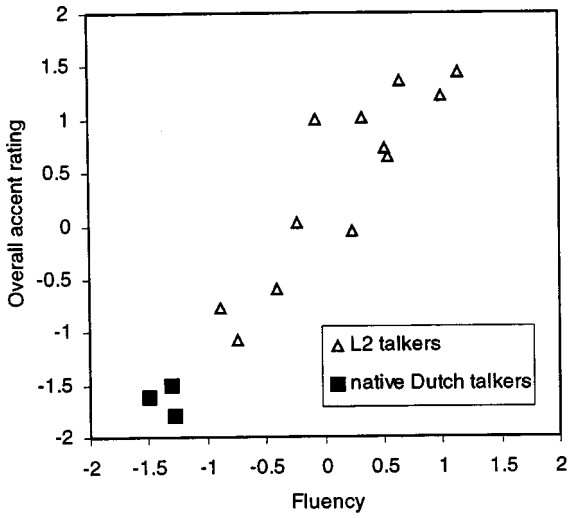


FIG. 4. Relation between pairwise comparison ratings for "fluency" and overall foreign accent ($R^2=0.89$).

talker who is unable to produce certain speech sounds may be judged to have the same degree of accent as a talker with near-perfect articulation, who however speaks very dysfluently. Yet, it is reasonable to expect differences in speech intelligibility between these two talkers.

To find out if the overall accent ratings can be separated into two dimensions ("clarity of articulation" and "fluency"), the pairwise comparison experiment was repeated with ten listeners. The subjects were first exposed to all talkers and asked to give overall accent ratings. After this, they were asked on which criteria they based their decision. All ten subjects mentioned (in their own words) clarity and fluency. A short discussion about the difference between these dimensions was held to verify the subjects' proper understanding of the difference. Next, the subjects were explicitly asked to compare the pairs of talkers, based on only one of these two dimensions at a time.

The same ten listeners compared all pairs of talkers twice on both dimensions, in consecutive experiments. In the break between these two sessions, the difference between clarity and fluency was again discussed. The relation between the scores from these experiments and the overall accent ratings from the original pairwise comparison experiment is given in Figs. 3 and 4.

Clearly, the holistically perceived foreign accent is related to clarity of articulation as well as fluency. The very high correlation between the overall ratings and the ratings for clarity of articulation indicate that clarity of articulation is the most important factor for the perception of overall accent strength.

Some of the 15 talkers are relatively similar in terms of the severity of their foreign accent. The data is arranged more conveniently by grouping the talkers into categories of accent strength. The 15 talkers were divided into four categories of accent strength based on the pairwise comparison ratings. This division into categories is given in Table II.

## III. INTELLIGIBILITY OF SPEECH IN NOISE FOR NON-NATIVE TALKERS

### A. Methods

We expect non-native speech production to be influenced by factors at segmental *and* supra-segmental level. When we wish to include all possible supra-segmental effects in our quantification of speech intelligibility, we must apply a type of speech intelligibility test that uses speech tokens consisting of multiple words. A suitable test method for this purpose is the speech reception threshold, or SRT (Plomp and Mimpen, 1979). Although developed as an audiological screening tool, the SRT method has proven to be useful for multi-lingual and cross-language speech communication research (van Wijngaarden, 2001b). Speech intelligibility can be thought of as the success that a talker and a listener have in transmitting linguistic information. By measuring the "success rate" (intelligibility) at the receiving end of the channel (the listener), the performance of the whole chain from talker to listener is measured. To evaluate the difference in intelligibility when L2 talkers are introduced, results are compared the results of baseline (L1) experiments.

A suitable method for investigating speech intelligibility at the phoneme level is the semi-open response consonant-vowel-consonant test (van Wijngaarden, 2001a).

#### *1. Subjects*

The same 15 talkers were used as in the accent rating experiment. A group of 20 Dutch university students of various disciplines (not including languages or phonetics), aged 17–26, were recruited as listeners.

#### *2. Speech reception threshold (SRT) method*

The SRT test gives a robust measure for sentence intelligibility in noise, corresponding to the speech-to-noise ratio that gives 50% correct responses for short redundant sentences.

TABLE II. Separation of talkers into four different categories of foreign accent strength, according to pairwise comparison ratings $r$.

| Accent strength | Category I | Category II | Category III | Category IV |
|---|---|---|---|---|
| Accent rating $r$ | $r \leqslant -1$ | $-1 < r \leqslant 0$ | $0 < r \leqslant 1$ | $r > 1$ |
| Talkers | DM-1 | EM-8 | EF-7 | GM-5 |
| | DM-2 | CF-13 | PM-10 | CF-14 |
| | DF-3 | GM-4 | PF-12 | PF-11 |
| | GF-6 | | EM-9 | CF-15 |

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 masking noise spectrum 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, and decides whether the response is correct.

The first sentence of each list of 13 sentences is initially presented at such a low SNR that is very likely to be unintelligible to the listener. This same sentence is repeated until it is responded correctly, the SNR going up in 4-dB steps. This is done to quickly converge to the 50% intelligibility threshold. The remaining 12 sentences are only presented once. 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. By taking the average speech-to-noise ratio over the last ten sentences (ignoring the first sentences of the list to eliminate initialization effects), the 50% sentence intelligibility threshold (SRT) is obtained.

#### 3. Measuring the slope of the psychometric function for sentence recognition in noise

SRT scores characterize the psychometric function of sentence intelligibility by a single value: the SNR for which 50% sentence recognition occurs. Since sentence intelligibility as a function of SNR is known to be a steep function, the 50% point gives sufficient information for many applications. However, most speech communication in real life takes place at speech-to-noise ratios corresponding to other intelligibility levels 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.

By modeling the psychometric function as a cumulative normal distribution (e.g., Versfeld *et al.*, 2000), we can fully describe it with two parameters: the mean (which is the SRT) and the standard deviation (or, equivalently, the slope around the mean). These two parameters were determined by first measuring the SRT (50% point) following the standard procedure, and next measuring percentages of correct responses for SNR values 2 and 4 dB above and below the SRT value (using five sentence lists altogether). The mean and the slope of the psychometric function (in % per dB) around the 50% point were estimated by fitting a cumulative normal distribution through these points (Gauss–Newton nonlinear fit).

Before the actual SRT tests and slope measurement tests, all conditions were verified to yield 85% to 100% sentence recognition in the *absence* of noise (i.e., the psychometric function was tested for showing ceiling effects). This is a necessary requirement for the distribution-fitting procedure to yield meaningful results.

#### 4. Semi-open response consonant-vowel-consonant method

A semi-open-response CVC (consonant-vowel-consonant) intelligibility test, specifically developed for the purpose of testing phoneme intelligibility with non-native subjects, was used for measuring speech intelligibility at the phoneme level (van Wijngaarden, 2001a). Using nonsense consonant-vowel-consonant words, the recognition of 17 initial consonants and 15 vowels was systematically measured with 16 native listeners.

Because of the time-consuming nature of the test, only the three Polish talkers (PM-10, PF-11, and PF-12) were included, as well as a single native Dutch talker (DM-2) to serve as a native baseline. To measure the effect of noise on phoneme recognition, the experiments were carried out at four speech-to-noise ratios ($-9$, $-3$, $+3$, and $+9$ dB). The masking noise used in this experiment had a long-term spectrum equal to that of speech by the tested talker.

### B. Results and discussion

#### 1. SRT scores of non-native talkers

Speech reception thresholds for each of the 12 L2 talkers, as measured with 20 native listeners, were all equal to or higher than for the three native talkers. This means that the intelligibility of the L2 talkers is, as expected, equal or lower, compared to native speakers of the Dutch language. The mean SRT score for each talker is given in Table III.

The relation between perceived foreign accent and speech intelligibility is shown in Fig. 5.

Although there is a relatively high correlation ($R^2 = 0.70$), there is some residual variance in SRT scores that cannot be explained from foreign accent strength. This is partly normal inter speaker variability, which is also ob-

TABLE III. Mean SRT scores and associated standard errors ($N = 20$).

| Talker | Native language | Mean SRT | Standard error |
|---|---|---|---|
| DM-1 | Dutch | −0.22 | 0.29 |
| DM-2 | Dutch | −1.28 | 0.25 |
| DF-3 | Dutch | −1.12 | 0.26 |
| GM-4 | German | 2.5 | 0.39 |
| GM-5 | German | 2.7 | 0.32 |
| GF-6 | German | −0.46 | 0.26 |
| EF-7 | Am. English | 0.8 | 0.32 |
| EM-8 | Am. English | 0.38 | 0.24 |
| EM-9 | Am. English | 1.86 | 0.38 |
| PM-10 | Polish | 1.96 | 0.46 |
| PF-11 | Polish | 3.6 | 0.45 |
| PF-12 | Polish | 1.9 | 0.41 |
| CF-13 | Chinese | 0.68 | 0.46 |
| CF-14 | Chinese | 1.9 | 0.46 |
| CF-15 | Chinese | 0.82 | 0.30 |

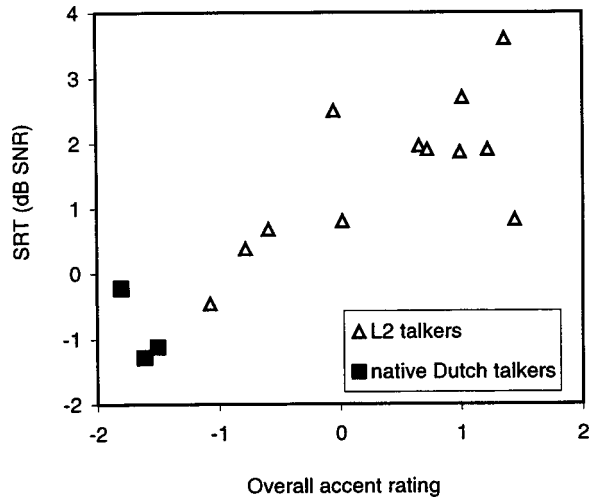


FIG. 5. Relation between foreign accent ratings and SRT scores for speech intelligibility. Accent strength is significantly correlated with speech intelligibility ($R^2=0.70$).

served for the native talkers. There is also a somewhat lower, but still significant, correlation between self-reported proficiency and SRT ($R^2=0.59$). This means that accent ratings from pairwise comparison experiments (Fig. 5) as well as self-ratings hold a predictive value for speech intelligibility.

When comparing Table III and Fig. 5 to similar data for non-native *listeners* instead of talkers (e.g., van Wijngaarden *et al.*, 2002), it appears that the effect of non-native speech production on intelligibility tends to be smaller than that of non-native perception. The worst-case SRT deficit for an L2 talker is around 5 dB in this experiment, compared to 7 dB for non-native listener of roughly comparable proficiency. In a within-subjects study comparing effects of L2 production versus L2 perception, perception was also found to be of greater influence (van Wijngaarden, 2001a).

### 2. Slope of the psychometric function for sentence reception

Because of the large number of test sentences needed per condition, the slope of the psychometric function for sentence recognition was not measured for all talkers, but only for one talker out of each category given in Table II. Since perceived accent strength and intelligibility correlate well, it can be assumed that the division into accent strength categories holds as a division in categories for intelligibility effects. The selected talkers are the ones closest to the mean of their category in terms of foreign accent rating.

An exception was made for native talkers; these were all three included, in order to be able to get an impression of the regular (native) interspeaker variability. The mean of the psychometric function and the slope around the 50%-point are given in Table IV.

Please note that the 50%-point of the psychometric function as reported in Table IV is essentially the same measure as the SRT reported in Table III, but determined with another paradigm. The correspondence between these values for the same talkers is good; the difference is smaller than 0.4 dB for any talker.

TABLE IV. Mean (SRT) and slope of the psychometric function for sentence recognition in noise. Means and standard errors across five listeners are given.

| Talker | Accent category | Native language | 50%-point (dB) | s.e. 50%-point | Slope around 50% (%/dB) | s.e. slope |
|---|---|---|---|---|---|---|
| DM-1 | I | Dutch | 0.2 | 0.3 | 12.2 | 1.0 |
| DM-2 | I | Dutch | −1.0 | 0.4 | 13.4 | 1.4 |
| DF-3 | I | Dutch | −0.7 | 0.4 | 12.2 | 1.2 |
| CF-13 | II | Chinese | 0.7 | 0.4 | 10.5 | 0.9 |
| PM-10 | III | Polish | 1.8 | 0.4 | 8.9 | 0.8 |
| PF-11 | IV | Polish | 3.6 | 1.1 | 8.3 | 1.5 |

Table IV shows that, as proficiency increases, the mean of the psychometric function shifts, but the curve becomes steeper as well. This is further indicated by Fig. 6, which shows the full psychometric functions according to the data in Table IV, assuming that these follow a cumulative normal distribution.

Figure 6 clearly shows that the reduction of intelligibility of non-native speech depends both on the proficiency of the talker and the speech-to-noise ratio. It is interesting to observe that the psychometric functions coincide near 0%, at a speech-to-noise ratio that is more or less the same for native and non-native talkers. Only as the speech-to-noise ratio rises, do differences between the talkers become apparent.

### 3. Phoneme recognition

So far, all presented speech intelligibility data was based on complete sentences. In all cases, near-perfect intelligibility of these sentences was found to occur in the absence of noise. Such good performance, despite the influence of foreign accents, is largely possible because of context effects. The recognition of individual speech sounds is much aided by word and sentence context.

A complication arises when comparing the influences of different foreign accents—the relation between the native language of the talker and the language that is spoken is likely to have an important influence on the patterns of phoneme confusions that occur. To prevent confounding of this

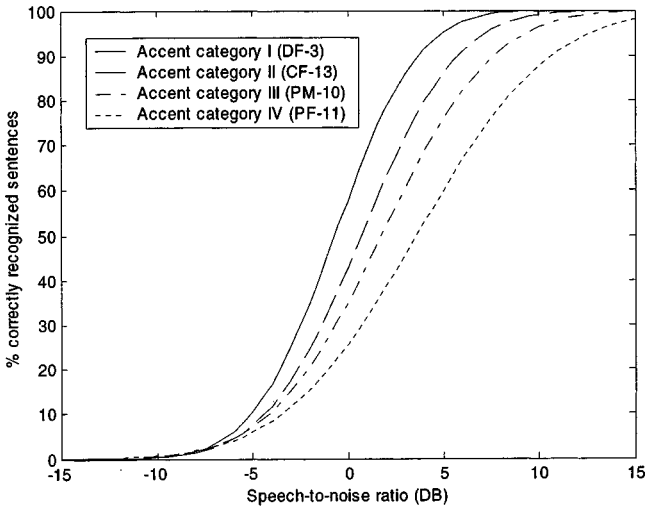


FIG. 6. Average psychometric functions for the recognition of sentences by four talkers, differing in accent strength.

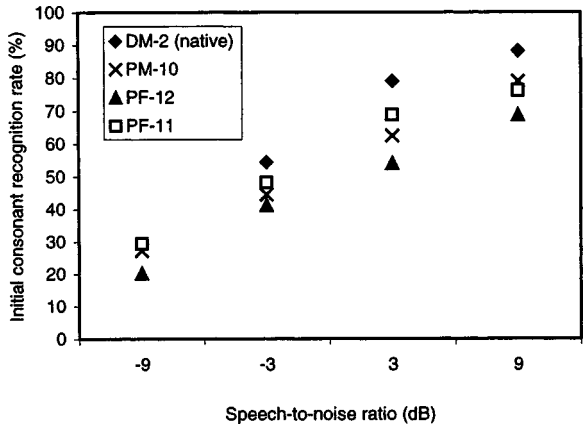


FIG. 7. Percentage of correctly recognized initial consonants in CVC words for three Polish and one Dutch talker speaking Dutch, as a function of speech-to-noise ratio (mean values across 16 native listeners; standard errors are in the range of 2–4.5 percent points).

effect with the effect of talker proficiency, only Polish talkers are compared to a (baseline) Dutch talker (see Figs. 7 and 8).

There is a clear (and statistically significant) overall effect of foreign accent on initial consonant recognition (Fig. 7), but the lowest-scoring talker is not the talker with the accent that was rated to be the strongest. At the highest speech-to-noise ratio (+9 dB), the ceiling for initial consonant recognition is not yet reached.

The recognition of individual vowels (Fig. 8) appears to be explainable by means of foreign accent strength: the stronger the perceived foreign accent, the lower the ''ceiling'' to which the percentage of correctly recognized vowels rises as the noise level decreases. This suggests that the L2 talkers consistently mispronounce some vowels. Since the talkers are from the same language background, one might expect that they all have difficulties pronouncing the same vowels. The Polish vowel system has eight vowels, of which six (/Iieaou/) also occur in Dutch, and are included in the CVC test. Individual realizations of these vowels differs between Dutch and Polish, depending on context; specifically, vowel duration is used differently in Dutch than in Polish. Hence, these six vowels are in practice not always the *same*

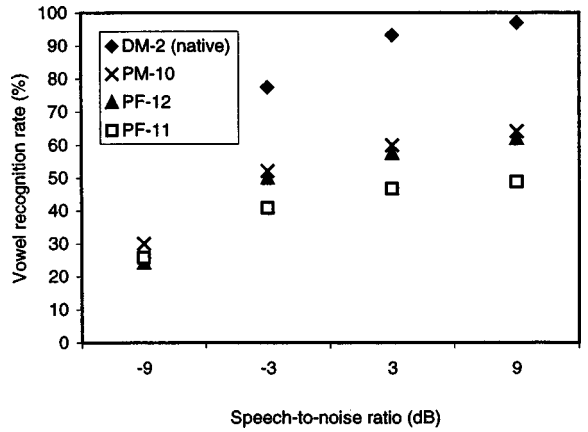


FIG. 8. Percentage of correctly recognized vowels in CVC words for three Polish and one Dutch talker speaking Dutch, as a function of speech-to-noise ratio (mean values across 16 native listeners; standard errors are in the range of 2.2–5.3 percent points).

TABLE V. Values of $R^2$ (explained variance) from an analysis of the correlation between specific vowel recognition errors for individual talkers. High values of $R^2$ indicate that the recognition errors of the 15 individual vowels follow the same patterns for each of the individual talkers. None of the correlations is statistically significant.

| $R^2$ | DM-1 | PM-10 | PF-12 | PF-11 |
|---|---|---|---|---|
| DM-1 | ··· | 0.17 | 0.03 | 0.01 |
| PM-10 | 0.17 | ··· | 0.06 | 0.07 |
| PF-12 | 0.03 | 0.06 | ··· | 0.01 |
| PF-11 | 0.01 | 0.07 | 0.01 | ··· |

in both languages, but are always at least *similar*. The other nine vowels included in the Dutch CVC test (including three diphthongs) do not occur in Polish at all.

To see if the patterns of vowel confusions are consistent across talkers, the percentage of correct recognition is calculated separately for each of the 15 tested vowels. The correlation between these specific vowel recognition scores indicates whether or not the vowel confusion patterns are consistent between L2 talkers. As Table V shows, there seems to be no consistency, despite the common language background of the L2 talkers. This was also informally observed by inspecting vowel confusion matrices for the individual talkers. The lack of consistency in auditory judgments of L2 speech sounds is a known phenomenon (Leather, 1983). When testing hypotheses regarding the L2 speech learning process, this inconsistency is experienced as a practical problem. However, when quantifying the intelligibility of cross-language speech communication, it reflects the situation that occurs in practice: poorly pronounced speech sounds are less likely to be correctly heard, but what they *will* sound like to the listener is unpredictable.

The speech learning model (SLM; Flege, 1992, 1995) predicts that late L2 learners, such as the Polish talkers in our experiments, are able to master *completely new* L2 sounds to perfection, if provided with sufficient phonetic input. Speech sounds that are *similar* to sounds that occur in L1 are never completely learned; these sounds are ''mapped'' onto L1 categories in L2 perception and production. For our CVC experiment, this implies that we may expect different relations between overall proficiency and recognition of the nine new versus the six similar vowels. In Fig. 9, the scores for ''new'' and ''similar'' vowels are given for the different talkers.

The recognition of new vowels does not differ significantly between the L2 talkers, despite differences in proficiency and overall intelligibility. The recognition of similar vowels *does* differ between L2 talkers: the lowest-proficiency talker shows the lowest overall recognition percentage of vowels that are similar to Polish vowels. For this talker (PF-11), new vowels are recognized better than similar vowels, while for talker PF-12 the opposite is true. When regarding the proficiency difference between PF-11 and PF-12, the difference in vowel recognition patterns is as predicted from Flege's SLM (Flege, 1995).

### 4. Relation between phoneme and sentence intelligibility

The overall recognition of sentences (Fig. 6), although fundamentally based on phoneme recognition, follows a

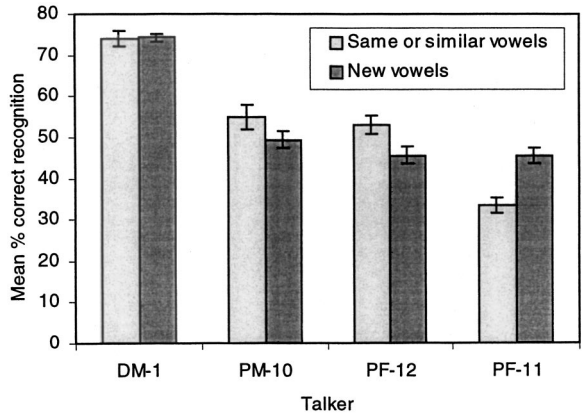


FIG. 9. Percentage of correctly recognized vowels for two sets of vowels: Dutch vowels that are the same (or similar) in Polish, and Dutch vowels that are new to Polish learners of the Dutch language. The error bars indicate the standard error ($N=16$; mean percentages taken per listener).

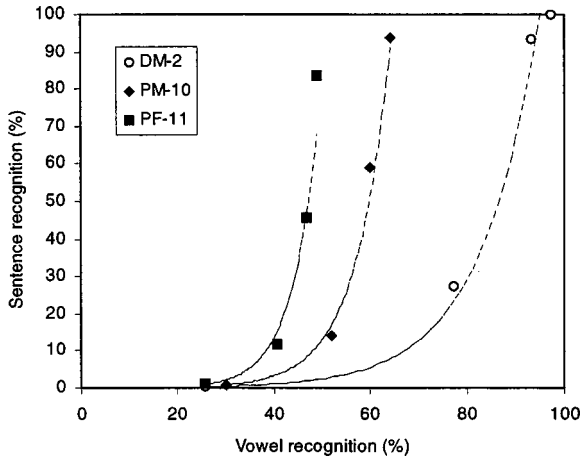


FIG. 10. Sentence recognition as a function of vowel recognition for three talkers: one native, two non-native) at four different speech-to-noise ratios ($-9$, $-3$, 3, and 9 dB). To guide the eye, and exponential curve is fit to the data of each talker.

somewhat different pattern than the recognition of individual phonemes (Figs. 7 and 8). The difference that is perhaps noted first is that ceiling effects as observed for vowel recognition appear absent from sentence recognition results.[1] When no noise is present, the sentences are sufficiently redundant to allow native listeners to make up for the faulty recognition of individual phonemes by making use of sentence context.

For native speech, when assessing speech intelligibility in rooms, or speech transmission quality of communication channels, the applied methods mostly make use of phoneme-level stimuli. Although one is invariably interested in transmission of complete messages rather than individual phonemes, there are good reasons to use a phoneme-based method. An advantage over sentence-based tests is that phoneme tests do not have such a steep transition around 50%, giving a better coverage of the range from excellent to very poor conditions. As long as a one-to-one relation between phoneme and sentence intelligibility is observed, phoneme intelligibility can be used as a predictor for the intelligibility of entire messages. Ceiling effects do, in this case, occur for vowels (Fig. 8), and perhaps also for consonants. This means that this condition is apparently not always met for non-native speech; hence, phoneme-based results can not always be relied upon as a predictor for the intelligibility of messages. This is further illustrated by Fig. 10, which combines data from Figs. 6 and 8.

Because of the ceiling effects in the vowel recognition scores, the (nearly) one-to-one relation between sentence and vowel intelligibility observed for the native talker is not realized for the non-native talkers. This does not mean that the intelligibility of non-native speech can *never* be predicted from phoneme level results. In this case, for instance, initial consonant recognition can be used to predict sentence intelligibility much better than vowel recognition. However, the current results indicate that phoneme-based measures that are known to predict sentence intelligibility in native speech require validation before applying those measures to non-native speech.

## IV. RELATION BETWEEN SPEECH INTELLIGIBILITY AND ACOUSTIC-PHONETIC MEASURES

### A. Global acoustic measures

The effects of specific talker-related influences on speech intelligibility are clearly present in the speech signal, since these are related to the source of this speech: the non-native talker. It is thus conceivable that an acoustic-phonetic analysis of foreign accented speech could yield objective predictions of the effect of foreign accent on intelligibility. The potentials of having such objective predictions, if sufficiently reliable, are great. Instead of needing a talker's self-ratings for foreign accent, or some other measure that may be difficult to obtain, intelligibility can then be predicted from physical measurements. Within the scope of this article, only relatively simple acoustic-phonetic measures were considered, because methods that are complex or require great annotation effort will probably have limited applicability.

Bradlow *et al.* (1996) distinguish ''global'' and ''fine-grained'' talker characteristics in predicting the influence of acoustic talker characteristics on speech intelligibility. Typical global characteristics are measures related to pitch and speaking rate; typical fine-grained characteristics include phoneme categorization and segmental timing relations.

To investigate the relation between speaking rate and intelligibility for non-native talkers, the results from the SRT experiments (Table III) were used. The SRT sentence recordings had been paced by means of a visual time indicator, allowing the talkers up to a maximum of 2.6 s for each SRT sentence. The talkers had been instructed to maintain a constant speaking rate across all sentences, trying to use as much of the 2.6-s ''recording window'' as possible. Despite the use of this pacing method, small (and in some cases statistically significant) differences in talking rate were observed between talkers (0.40–0.65 sentences per second). An analysis of the relation between talking rate and SRT revealed, however, no significant correlations. Mean $F0$ and $F0$ range (mean difference between highest and lowest $F0$ in a sentence) were found to vary across talkers, to the same degree for native as well as non-native talkers. The latter indicates

TABLE VI. Percentage of vowels (ten vowels, three realizations each) correctly classified according to the vowel regions by Pols *et al.* (1973).

| Talker | Accent category | Correctly classified (%) |
|---|---|---|
| DM-1 (native) | I | 83.3 |
| DM-2 (native) | I | 86.7 |
| GM-4 | II | 60.0 |
| EM-8 | II | 70.0 |
| EM-9 | III | 63.3 |
| PM-10 | III | 63.3 |
| GM-5 | IV | 73.3 |

that pitch variations are applied by native *and* non-native talkers. However, $F0$ and $F0$ range did not correlate significantly with SRT or CVC results, meaning that these measures can not be used as predictors of speech intelligibility.

### B. Fine-grained acoustic measures

A more fine-grained talker characteristic that is known, at least for native talkers, to correlate with speech intelligibility, is vowel space size (e.g., Bradlow *et al.*, 1996). Larger vowel spaces tend to lead to more intelligible speech in native talkers.

Of each of the 15 talkers, mid-vowel formant frequencies were calculated for 3 stressed instances of 11 different Dutch vowels. First, the overall variance in $F1$ and $F2$, for all 33 vowels of each talker, was considered, as a broad estimate of vowel space size. This variance did not correlate with SRT results ($R^2=0.03$, across 15 talkers), nor with CVC vowel recognition scores ($R^2=0.07$, across 4 talkers). This means that the size of the vowel space does not predict intelligibility differences between non-native talkers.

The ratio between within-vowel variance and overall variance was also determined. In this way, essentially by comparing the statistical spread of different instances of the same vowel to the spread of *all* vowels, a coarse indication of ''discriminability'' in the $F1$-$F2$ plane is obtained. However, this variance ratio does not correlate significantly with non-native CVC or SRT results either.

For non-native talkers, one could expect the decreased intelligibility to result from a distorted rather than just a reduced vowel space. Distortion, in this context, is not as easy to measure as reduction, since it requires a priori knowledge of how the vowel space should be organized to be perceptually acceptable. Such a priori knowledge can in some cases be taken from vowel space studies, such as reported by Pols *et al.* (1973) for Dutch vowels of 50 male talkers. Pols *et al.* defined vowel categories in the $F1$-$F2$ plane as maximum-likelihood regions, indicating clear borders between categories. The same $F1$-$F2$ data as used for calculation of the variance ratios was applied to determine which percentage of the vowels are correctly categorized according to the regions by Pols *et al.* (only for the male talkers). The results are given in Table VI.

The scores for the two male native talkers are higher than for the non-native talkers. The mean percentages of correct classification per vowel, for all of the talkers in Table VI, were subjected to a two-way ANOVA (the two factors being native/non-native and vowel category). A significant ($P<0.01$) main effect of native versus non-native was found. The percentage correct classification was also found to correlate significantly with accent ratings ($R^2=0.57$) and SRT ($R^2=0.67$). This means that of the acoustic-phonetic measures that were considered in this study, this is the only one that was found capable of predicting intelligibility effects of non-native speech. Unfortunately, it is also the measure that is the most difficult to obtain. It requires detailed and reliable a priori knowledge of the native $F1$-$F2$ plane, and hand-labeling of suitable stressed vowels for each talker.

## V. GENERAL DISCUSSION AND CONCLUSIONS

Foreign accented speech tends to be less intelligible than native speech. The results presented in this article confirm that L2 experience is an important determining factor for the intelligibility of a non-native talker.

The overall effect on speech intelligibility is proportional to the degree of foreign accent ($R^2=0.70$). Hence, by estimating the severity of a talker's accent, a first impression of the intelligibility effects is obtained. Moreover, a talker's own opinion of his L2 proficiency can also be used as a predictor of speech intelligibility ($R^2=0.59$).

For non-native speech, the recognition of individual phonemes may sometimes be impaired even in the absence of noise. In the case of the Polish subjects who participated in this study, this was found to be the case for a large fraction of the Dutch vowels. Nevertheless, sentence intelligibility could still reach 100%. This shows the powerful effect of contextual information in human speech recognition. The practical implication for quantifying the overall effects of foreign accent on speech intelligibility is that sentence-based methods seem to be more suitable than phoneme-level methods. Before using any phoneme-level test result to predict the intelligibility of non-native speech, the existence of a reversible one-to-one relation needs to be established.

Objective phonetic-acoustic measurements are not easily applied to predict effects of foreign accent on intelligibility. Of several global and fine-grained acoustic phonetic measures, the only one found to correlate significantly with intelligibility was a measure that quantifies the deviations between a talker's own (non-native) vowel realizations to the native $F1$-$F2$ plane. However, this measure is not particularly suitable for intelligibility predictions. The fact that the process of obtaining this measure is laborious, and requires detailed knowledge of the native $F1$-$F2$ plane, was already mentioned. Moreover, the measure is only concerned with vowels. The relation between vowel recognition and sentence intelligibility was shown *not* to be a one-to-one relation for non-native speech; any measure related to vowel space should be expected to suffer the same limitations.

As a final note, it is important to realize that all experiments described in this article were concerned with the intelligibility of *recorded* non-native speech. In real conversations, non-native talkers have the ability to respond to listeners' apparent comprehension of their speech. They are also less likely to use words or grammatical constructions

they are not familiar with, which may very well lead to a better overall speech intelligibility.

## ACKNOWLEDGMENTS

The authors would like to thank Adelbert Bronkhorst for his helpful comments on an earlier version of this manuscript.

[1]One could argue that the psychometric functions of Fig. 6 are the result of modeling the psychometric function as a cumulative normal distribution, and will therefore always go up to 100%. However, the individual responses on which the calculation of the psychometric function is based show that saturation at 100% (or very close to 100%) is in fact observed for native as well as non-native speech.

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.

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.

Flege, J. E. (**1984**). ''The detection of French accent by American listeners,'' J. Acoust. Soc. Am. **76**, 692–707.

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).

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.

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

Leather, J. (**1983**). ''Second-language pronunciation learning and teaching,'' Language Teach. **16**, 198–219.

Magen, H. S. (**1998**). ''The perception of foreign accented speech,'' J. Phonetics **26**, 381–400.

Peterson, G. E., and Barney, H. L. (**1952**). ''Control methods used in a study of the vowels,'' J. Acoust. Soc. Am. **24**, 175–184.

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

Pols, L. C. W., Tromp, H. C. R., and Plomp, R. (**1973**). ''Frequency analysis of Dutch vowels from 50 male speakers,'' J. Acoust. Soc. Am. **53**, 1093–1101.

Singh, S. (**1966**). ''Crosslanguage study of perceptual confusion of plosive phonemes in two conditions of distortion,'' J. Acoust. Soc. Am. **40**, 635–656.

Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. (**1994**). ''Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude,'' J. Acoust. Soc. Am. **96**, 1314–1324.

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

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

van Wijngaarden, S. J. (**2001b**). ''Methods and models for the assessment of cross-language speech intelligibility,'' Workshop on Multi-lingual Speech and Language Processing, Aalborg, Denmark, 8 September.

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