# The intelligibility of non-native speech

Sander J. van Wijngaarden

VRIJE UNIVERSITEIT

# THE INTELLIGIBILITY OF NON-NATIVE SPEECH

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geneeskunde
op woensdag 8 oktober 2003 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Sander Jeroen van Wijngaarden

geboren te Rotterdam

promotor: prof.dr.ir. T. Houtgast
copromotor: dr.ing. H.J.M. Steeneken

# THE INTELLIGIBILITY OF NON-NATIVE SPEECH

Sander J. van Wijngaarden

Het in dit proefschrift beschreven onderzoek werd uitgevoerd bij
TNO Technische Menskunde in Soesterberg.

The research reported in this dissertation was carried out at
TNO Human Factors, Soesterberg, the Netherlands.

**The intelligibility of non-native speech**

Sander Jeroen van Wijngaarden
Ph.D. thesis, Amsterdam, 2003
Keywords: speech intelligibility, non-native speech, cross-language

The cover design is based on translations of the first sentence of the
original Speech Reception Threshold test (Plomp and Mimpen, 1979)

Printed by PrintPartners Ipskamp BV, Enschede

ISBN 90-76702-03-9

*Aan Judith, Onno en mijn ouders*

# Voorwoord

Ik heb nooit zoveel gehad met voorwoorden.

Met uitzondering van proefschriften sla ik een voorwoord bijna altijd ongelezen over. Het voorwoord van een proefschrift is echter iets uitzonderlijks: de promovendus laat zien dat hij, ondanks het feit dat hij apetrots is op zijn versgedrukte geesteskind, belangrijke waarden zoals dankbaarheid en nederigheid niet uit het oog heeft verloren ("ik had het nooit gekund zonder ..."). Men kan zich daarbij redden met een eenvoudig basisrecept: "men neme zes willekeurige proefschriften, en menge de in voorwoorden aangetroffen clichés totdat een gladde, homogene massa ontstaat."

Ik zal mijn best doen om aan de verwachting te voldoen, daarbij de gemeenplaatsen zoveel mogelijk vermijdend.

Ik ben in de eerste plaats dank verschuldigd aan twee personen: Tammo Houtgast en Herman Steeneken. Zij hebben mij de kans gegeven om dit promotie-onderzoek uit te voeren binnen mijn TNO-dienstverband. Dit is een bevoorrechte positie, die ik iedereen die de kans krijgt van harte kan aanbevelen.

Herman Steeneken was vanaf het moment dat ik bij TNO kwam werken mijn mentor. Ik heb in Delft de basisbeginselen van de akoestiek geleerd, maar Herman heeft mij het akoestisch vakmanschap pas echt bijgebracht. Daarnaast heeft hij mij vanaf het allereerste begin alle kansen en vertrouwen gegeven. Dit heeft mij enorm gestimuleerd.

Tammo Houtgast is voor mij, en voor vele van mijn collega's, de ultieme wetenschapper. Er is niemand die feillozer de zwakke plek in een willekeurige redenering kan aanwijzen. Mijn wetenschappelijke vorming dank ik aan meerdere leermeesters, maar vooral aan Tammo.

Ik ben vele *reviewers*, zowel intern als extern, erkentelijk voor hun bijdragen aan de verbetering van mijn werk.

Adelbert Bronkhorst heeft mij, destijds als groepsleider verantwoordelijk voor het goedkeuren van mijn manuscript voor hoofdstuk 3, enkele rake klappen uitgedeeld. Mede dankzij hem is dit proefschrift tientallen pagina's korter uitgevallen en ontbreken enkele moeilijk verdedigbare conclusies. Wellicht was mijn verschrijving in het

*acknowledgement* bij een manuscript-versie van een JASA-paper Freudiaans ("the *others* would like to thank Adelbert Bronkhorst..." in plaats van *authors*). In elk geval ben ik inmiddels alle frustratie voorbij en resteert dankbaarheid voor de resulterende kwaliteitsverbetering.

I would also like to thank Søren Buus, who provided many valuable suggestions on Chapter 4 when it was originally submitted to JASA, and on the whole dissertation in a later stage.

Verder ben ik dank verschuldigd aan alle leden van de leescommissie voor hun nauwgezette commentaar. Ik wil daarbij met name Bert Schouten noemen, die dit proefschrift tevens heeft ontdaan van een indrukwekkend aantal (kleine en minder kleine) taal- en spellingsfouten.

Natuurlijk ben ik ook vele collega's, zowel binnen de afdeling als daarbuiten, dankbaar voor hun steun, assistentie en kameraadschap. Ik zal slechts enkelen met name kunnen noemen. In elk geval spreek ik mijn welgemeende dank uit aan alle collega's van de groep Spraak & Gehoor; onder andere omdat zij drie jaar lang zonder morren geaccepteerd hebben dat ik een aanzienlijk deel van de beschikbare ruimte voor achtergrond-onderzoek in beslag nam.

Met David van Leeuwen heb ik, eerst als kamergenoot en later zonder enig alibi, vele uren aan vruchtbare discussies over mijn promotie-onderzoek besteed, in tijd die eigenlijk voor ander onderzoek bedoeld was. Ook wil ik in het bijzonder mijn latere kamergenoot Ronald noemen (dankbaar mikpunt van practical jokes), en natuurlijk Jan Verhave (duivels uitvinder van practical jokes). En laat ik ook Rob niet vergeten – die zelfs in Pittsburgh Wyoming Tennessee Ohio de pizza-outlet weet te vinden.

I would particularly like to thank all of my subjects, who gave their time and energy to participate in the various experiments reported in this dissertation. Of the 150+ subjects (of 10 nationalities) who took part, some were willing to return for quite a number of different sessions. Without these people (in particular Kyle, Sandi, Ela, Bernd, and Klara), some of my experiments would have been impossible to complete.

Vele vrienden, collega's en kennissen (en kennissen van vrienden van collega's) hebben geholpen met de vertaling van "de bal vloog over de schutting". Zij hebben het ontwerp voor de omslag van dit proefschrift mogelijk gemaakt, waarvoor mijn dank.

Tenslotte ben ik natuurlijk ook veel dank verschuldigd aan Judith, die ondanks haar eigen promotie-stress de tijd, moed en energie wist te vinden om mij te motiveren, en en-passant een prachtige zoon op de wereld te zetten. Met haar deel ik de dingen die belangrijker zijn dan welk proefschrift ook.

# Contents

# Chapter 1. General Introduction

Speech, often claimed to be the most important means of communication between humans, sounds profoundly different between different regions of the world. Experts seem to disagree on the number of languages still in use today; around 6000 is a popular estimate. A first indication, which can be given with relative certainty, is that approximately 2200 (partial) translations of the Bible alone are known (Kanungo and Resnik, 1999). With that many languages in use around the world, the probability of coming across a language barrier seems enormous.

Normally, language barriers only come into play when geographical distances are traversed, either through physical travel or by means of telecommunication technology. As it happens, the last few decades have brought us dramatic increases in global travel as well as telecommunication. As a consequence, almost everybody being a native talker of only a single language, the number of conversations that are impaired by "non-native" effects must be increasing.

In some situations, such as sound systems in public buildings, the intelligibility of spoken messages presented to the public is required by standards and regulations to exceed a certain minimum. In other cases, such as mobile telephony, speech intelligibility is a measure of the quality of service offered to the customer. For purposes such as these, methods have been developed that are capable of measuring and predicting speech intelligibility. Currently, these methods are usually based on the (implicit) assumption of fully native speech communication. An increasing need is felt to extend the scope of existing methods for measuring and predicting intelligibility to include populations of non-native communicators. This thesis hopes to answer a few questions that are encountered on the way to fulfilling this need, and perhaps to raise a number of new questions.

Whatever aspect of non-native speech communication is studied, the key question always turns out to be: what differences do we observe compared to native communication? When studying speech intelligibility, a natural approach is to compare measures of non-native intelligibility with a native baseline. An important requirement in the context of this study is also that the applied methods need to deliver quantitative results. It is often easy, if not trivial, to prove that a difference exists between native and non-native intelligibility. The challenge is to determine *to what degree*, and that with a

sufficient degree of accuracy. Only then can the results be applied in the context of intelligibility prediction models.

Another important consideration is *under what conditions* we are measuring non-native intelligibility. Clearly, native intelligibility is also largely determined by communication conditions; relations between physical measures of speech signal degradation (such as noise, reverberation and peak clipping) and intelligibility have been measured extensively. Non-native relations between, for instance, the speech-to-noise ratio and intelligibility, or reverberation time and intelligibility, should be expected to deviate from what is known for native communication. Will the difference be a simple shift of the psychometric function, or should we expect a more complicated relation between native and non-native intelligibility?

A third important variable is the *level of analysis* at which speech communication is studied. We have been getting used to finding more or less consistent relations between results found, say, on the level of individual phonemes, and on the level of words or sentences. In many cases, the motivation for measuring speech intelligibility lies in an interest in how well complete messages are exchanged (which normally consist of at least a number of words). Still, the intelligibility is often measured using phoneme-based tests, ignoring the influence of lexical, syntactic and semantic information. This is acceptable for native intelligibility, knowing that the predictive power of phoneme recognition for the intelligibility of messages is sufficient. For non-native intelligibility, this remains to be seen.

Two very important variables that are difficult to separate in experiments, are the *type of communication* and the *populations of communicators*. Is there a complex conversation going on, with a lot of interruptions and interaction between subjects, or is one talker simply addressing one listener? Is the listener non-native, or the talker, or perhaps both? What are the native languages of the subjects, and how are these related to the target language? What level of proficiency do the non-native subjects have with respect to the target language? Even when limiting the scope of a study to only a few languages, the number of configurations to be investigated can be enormous.

The experiments described in this thesis represent a cross section of the variable space described above. Choices were made; not everything that promised to be interesting was studied. The main motivation behind each choice was the potential importance for objective speech intelligibility prediction models, specifically the Speech Transmission Index (STI; Steeneken and Houtgast, 1980).

The STI takes a central position in the final chapters of this thesis. The main goal of this study is to adapt the STI for non-native speech communication. The standardized (native) STI method assumes a "standard" population of talkers and listeners. When the STI method is to be adapted for non-native use, it must also be supplied with information on the

populations of non-native communicators. Methods for describing communicators on dimensions that are relevant for non-native communication (broadly: proficiency) were sought that are easily measured, simple, and accessible.

All experiments are centered around the Dutch language. Having a single common language sometimes simplifies matters, creating (for instance) the possibility to use the same baseline data across various experiments. Given the fact that all experiments were carried out in the Netherlands, one can imagine the practical advantages of choosing Dutch as a central language. An interesting question is how the choice of languages affects the outcome of the study. Will a similar study, but centered around French, Danish or Italian, result in different findings? Although the proof is incomplete, there are convincing indications that the general trends and conclusions resulting from this study are largely language-independent. Where parts of this study lend themselves for comparison to studies centered around the English language (e.g., Florentine et al., 1984; Mayo et al., 1997), no important discrepancies are found.

Chapter 2 of this thesis goes into the selection of suitable test methods for cross-language measurements of speech intelligibility. It also outlines an approach to obtaining a sensible cross section through the vast space spanned by the relevant variables influencing non-native speech, as mentioned above.

Chapter 3 gives results on experiments in which the talker is the non-native factor. In Chapter 4, the experiments are centered around non-native listeners.

Chapter 5 combines some of the data given in earlier chapters, and shows a possible way to interpret the Speech Transmission Index when non-native speech is involved. In Chapter 6, a specific aspect influencing native as well as non-native speech is addressed: the *speaking style* adopted by talkers.

The final chapter summarizes all conclusions drawn throughout this thesis, and contains a general discussion on the approach that was adopted.

# Chapter 2. Methods and models for quantitative assessment of speech intelligibility in cross-language communication[1]


## ABSTRACT

To deal with the effects of non-native speech communication on speech intelligibility, one must know the magnitude of these effects. To measure this magnitude, suitable test methods must be available. Many of the methods used in cross-language speech communication research are not very suitable for this, since these methods are designed to investigate specific effects regarding speech perception and production, rather than quantifying overall intelligibility. In this chapter, a simple descriptive model of cross-language speech intelligibility is shown that helps in selecting experimental methods to assess speech intelligibility. Based on this model, and on practical observations regarding assessment of cross-language speech intelligibility, a multi-lingual version of the Speech Reception Threshold method was implemented as a suitable method for the quantification of cross-language speech intelligibility.


## 2.1. INTRODUCTION

Most reported experiments concerning non-native speech intelligibility have been designed to obtain a better insight into the details of the speech perception and production process. Researchers in the field of second-language speech production and perception usually aim to test very specific hypotheses. Which experimental method is the most efficient depends on the hypothesis tested.

Apart from research on the basics of human speech communication, an increasing need is felt for a more application-oriented approach, aiming at the overall effect on speech intelligibility. Cross-language speech communication,

[1] Modified version of a previously published paper: 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 *Proc. RTO Workshop on Multi-lingual Speech and Language Processing*, Aalborg, Denmark.

in which one or more parties engaged in a conversation depend on second-language skills, is an increasingly common phenomenon. The efficiency of cross-language speech communication is quite often experienced to be lower than 'fully native' communication. For many of those situations, it would be helpful to be able to assess the magnitude of the effect on speech intelligibility. Applications that could benefit from such knowledge would be, for example, the design of public address and communications systems, and prediction models in room acoustics. By knowing the extent to which speech intelligibility is reduced, better design criteria can be established.

Wanting to know the *extent* to which speech intelligibility is influenced means that quantitative methods for measuring speech intelligibility are needed. This is different from the hypothesis-driven methodology preferred for investigating the principles of non-native speech communication; instead of looking for *causes,* we are quantifying the *consequences.*

To illustrate this approach, consider the following situation. Suppose that an auditorium in a Dutch school is equipped with an air-conditioning system, which produces a known level of background noise. In 'normal' (native) situations, the intelligibility of the public address system in the auditorium is generally acceptable, despite the background noise. What if a native English talker addresses the Dutch students (in English), who have an average experience with the English language of 2 years? What if the average experience of the students is 5 years, or what if the native language of the talker is German? What reduction of the background noise level is necessary to obtain a certain minimum speech intelligibility?

When using suitable methods, it is possible to answer all of these questions, if populations of talkers and listeners are properly defined. Not all the *causes* behind the differences in intelligibility have to be known. These causes may be very complex, involving better analysis of the speech signal into phonetic units, larger vocabulary, better understanding of the grammar, etc. Regardless of the causes, the size of the effects is important in its own right.

In this chapter, we will present a simplified model of non-native speech communication. The aim of this model is to serve as a tool that helps in choosing the proper methods to quantify the effects on intelligibility. Based on this model, we will describe a multi-lingual speech intelligibility evaluation method that is suitable for application to cross-language speech communication.

## 2.2. A MODEL OF CROSS-LANGUAGE SPEECH COMMUNICATION

### 2.2.1. Types of cross-language speech communication

Describing a specific cross-language conversation unambiguously takes a little consideration. As the number of people engaged in a conversation increase, the complexity of a proper description of the situation increases accordingly.

All situations can be broken down into variants of straightforward two-way communication, in which case only one person is talking, and only one other person is listening. This involves influences from up to three languages: the native language of the talker, the native language of the listener, and the language that is currently being spoken. The relations between these three languages will partly determine the speech communication process. Comparative studies of the languages involved could theoretically shed light on this; analyses of phonetic contrasts and inspection of the (sound-based) lexicon of a specific language could help to understand its relation with other languages, provided this same information is also known for these other languages. Rather than trying to find a general model for language-related influences on cross-language communication, we will treat each combination of languages as a unique case.

It has become conventional to denote native talkers and listeners as 'L1,' and non-native (second-language) talkers and listeners as 'L2'. Based on this notation, one could (for example) indicate that a native listener is listening to a non-native talker by writing 'L2>L1'. This notation works if the number of languages involved is no more than two. The situation 'L2>L2' could mean that a Dutch listener is speaking English to a German listener; it could also mean that a Dutch listener is speaking English to another Dutch listener. The difference may be important, since the common native language between talker and listener may influence their use of the second language (in our example English).

To avoid confusion, we will use the following notation throughout this thesis:

Dutch > (English) > German

meaning that a Dutch talker is talking English to a German listener. We will generally abbreviate this to D>(E)>G.

### 2.2.2. Defining populations of talkers and listeners

Considering non-native speech intelligibility separately for each individual that comes our way would become a very laborious process. By defining meaningful populations of talkers and listeners, we can collect more generally applicable quantitative results. First, we decide what populations we need to

have quantitative data on; then we recruit subjects from these populations, and carry out experiments. Experiments may involve subjects selected from one single population, or may use talkers from one population and listeners from another.

In order to define a population, one should be able to describe it in terms of the determining factors for non-native speech intelligibility. The description of the population starts with the native language of the subjects; preferably, details concerning regional accents (if any) should also be known.

A very important factor is the average experience of subjects within the population with the target (second) language (e.g., Flege, 1992; Strange, 1995). Age of acquisition of the second language is also of great importance (e.g., Flege, 1995; Flege et al., 1997; Mayo et al., 1997).

Second-language experience and age of acquisition combine into second language *proficiency*, a term we will use rather loosely to indicate the underlying dimension explaining differences in non-native speech intelligibility. Despite the fact that second-language proficiency comprises many different abilities (related to phonetic discrimination, vocabulary, grammar, etc.), subjects are able to rate their own proficiency with a sometimes impressive accuracy (van Wijngaarden et al., 2001b; also see Chapter 3).

Possible other factors to consider could be more general descriptors of the population, such as age and gender. It seems fair to consider the influence of these variables on cross-language communication to be higher-order effects, but it seems only prudent to keep variables like these in mind as well when selecting subjects for experiments.

Even when the populations of talkers and listeners are fully defined, the resulting speech intelligibility may still vary according to numerous other variables, most of which also apply to fully *native* communication, such as speaking rate and speaking style. These variables are not really related to the characteristics of the talkers and listeners, but rather to their *mode* of communication. One aspect related to this is worth mentioning. For non-native talkers, the distinction between *read* speech and *spontaneous* speech is potentially of far greater importance than for native talkers. Non-native talkers are likely to limit their effective vocabulary to easier and more familiar words when speaking spontaneously, while they are more likely to produce pronunciation errors when asked to read a certain text aloud. In the latter case, they are not only likely to mispronounce unfamiliar words, but a poor understanding of context may also lead to an impaired intonation of sentences.

### 2.2.3. Conditions for speech communication

Native as well as non-native speech can be affected by adverse conditions, such as background babble, ambient noise, bandwidth limiting, or reverberation. However, the degrading influence on cross-language speech

communication tends to be greater (Gat and Keith, 1978; Lane, 1963; Mayo et al., 1997; Nábelek and Donahue, 1984; van Wijngaarden, 2001).

Measuring speech intelligibility under clear, undegraded, conditions is often not very effective. The effects of non-nativeness on intelligibility may be relatively small, whereas problems are to be expected in practice, when degrading circumstances *are* normally present. By conducting experiments under conditions that represent a controlled degree of speech signal degradation, the effect of this degradation on cross-language speech communication may be assessed systematically.

Perhaps the easiest way to reduce speech intelligibility in a controlled manner, is by adding stationary noise with a known spectrum. For fully native speech communication, intelligibility in this case is a relatively stable and well-known function of the speech-to-noise ratio. For non-native speech communication similar relations are found (van Wijngaarden, 2001; van Wijngaarden and Steeneken, 2000), which clearly show that noise is capable of affecting cross-language communication more profoundly than native speech communication.

### 2.2.4. Levels of analysis

Our approach towards the assessment of non-native speech intelligibility requires a model that describes cross-language speech communication in such a way, that the proper characteristics for quantifying intelligibility can be chosen.

In practice, this means that a description is needed of the determining factors for speech intelligibility (which we will call intelligibility cues), and an indication of where to find these. More specifically, we need to find out about intelligibility cues that are especially important when considering *cross-language* speech communication.

Speech intelligibility can be studied at various levels of analysis; the most basic analysis would involve studying the speech signal on an allophone-by-allophone basis. Perhaps the highest conceivable level would be to consider an entire story, where the amount of relevant information in the story that was transferred could be studied.

There are reasons to assume that the level of individual words takes an important position in the process of learning a second language (Bradlow and Pisoni, 1999); it seems likely that one initially learns a second language mainly by collecting a sound-based representation of its lexicon. For this reason, and because of practical considerations, we will distinguish three levels of analysis: speech units smaller than words (allophones), words, and speech units larger than words (sentences).

Besides the level of analysis, intelligibility cues can also be separated depending on whether they can be found in the speech signal ('acoustic' cues) or somewhere else. As an example of the difference: the intelligibility of sentences (as compared to the intelligibility of the individual words of which

they consist) is enhanced by means of intonation. Intonation (or more generally, prosody) is present in the speech signal, and can therefore be called an 'acoustic' intelligibility-enhancing factor. The semantic and syntactic redundancy contained in a sentence also increases its intelligibility relative to the individual words of which it consists. However, these factors can not be traced back to the speech signal; they improve intelligibility by aiding the listener in his cognitive processing of the message.

Table 2.1 illustrates the distinction between acoustic and non-acoustic intelligibility cues at the three defined levels of analysis.

Table 2.1. Levels of analysis in non-native speech communication

| Level of analysis | Examples of affected intelligibility cues | |
|---|---|---|
| | Acoustic | Non-acoustic |
| Supra-word level (sentence level) | Prosody | Syntactic constraints<br>Semantic constraints |
| Word level | Lexical contrasts | Word familiarity |
| Sub-word level (allophone level) | Phoneme inventory | |

This distinction between acoustic and non-acoustic factors is not helpful at the sub-word level. For the non-acoustic factors at this level (such as the individual phoneme space representation that a listener uses to categorize L2 allophones) can hardly be tested without involving acoustic allophone realizations.

Table 2.1 can be used to decide which *characteristic* of cross-language speech intelligibility is the most appropriate in a specific case, for instance phoneme recognition versus sentence intelligibility. Only after deciding what is the most appropriate characteristic can we design a proper experiment.

For example, one may wish to quantify the intelligibility of a group of (non-native) German actors, playing before an audience of native English listeners, in the English language (G>(E)>E). The non-acoustic intelligibility cues do not require special attention in this case, since only the talkers are non-native, and their vocabulary and sentence construction are 'programmed' by the play they are acting. Hence, all deviations from fully native communications can be found in the speech signal. At the very least, one may expect that the actors' allophone realizations will deviate from native English speech. A phoneme-based intelligibility test will be a suitable choice to quantify this effect. However, this may not be the *most* suitable intelligibility test. Unless the actors are thoroughly trained by a native English director or language coach, their intonation will also deviate from the

authentic English patterns. In that case, a (sentence-based) intelligibility test that is sensitive to differences in prosody is a better choice.

As another example, consider the reverse situation (the actors are now English and the audience is German; E>(E)>G). Since the German audience is now the only non-native factor, the speech signal is not at all affected. Still, the resulting speech intelligibility may be reduced considerably; partly because the non-native listeners are not as good at identifying individual speech sounds, but also for reasons related to vocabulary and the less effective use of word context (van Wijngaarden and Steeneken, 2000; also see Chapter 4). In this case, the average L2 linguistic development of the German audience is an important variable. Besides a speech intelligibility test using sentences (to include the effects of word context), it may be useful to include a separate test to quantify vocabulary and context-effects separately.

## 2.3. CONSIDERATIONS FOR SELECTING SPEECH INTELLIGIBILITY ASSESSMENT METHODS

### 2.3.1. Practical considerations

A pragmatic approach toward measuring non-native speech intelligibility is simply to adopt one of many proven experimental methods designed for native speech. Inevitably, some modifications to these proven methods will be necessary, if only for practical reasons.

Several intelligibility test methods are based on one-syllable nonsense words. These tests are generally quite efficient at measuring speech intelligibility at the phoneme level. Subjects participating in such tests must somehow communicate perceived nonsense syllables in response to the auditory stimuli. With L2 listeners, typing these responses should be ruled out as an option. Differences in orthographic representation of sounds between L1 and L2 will confuse the subject. Even highly proficient subjects, who are aware of differences in orthography between L1 and L2, are likely to produce errors, especially when working under time pressure. Collecting multiple-choice responses will partly solve this problem, especially if no 'confusing' alternatives are presented. In any case, proper subject instruction with regard to this issue is vital.

Some additional complications surrounding experiments with non-natives have to do with the recruiting of subjects. The definition of the population from which to draw subjects is much narrower than usual in speech intelligibility testing. Accordingly, subjects will be harder to find. Experimental methods can be designed or adapted to help cope with this issue. Methods that require special sound-insulated rooms or heavy equipment require subjects to travel to a certain location. By adapting these methods so that they can be implemented in a portable device (such as a

notebook computer), hard-to-reach subjects (unwilling to travel in order to take part in a test) can be tested at remote locations.

The available time per subject may also be shortened. When tests run over longer periods of time, a smaller percentage of the population of potential subjects will be willing to participate. By shortening the duration of the experiment (by making tests more efficient, or by spreading the load over a slightly larger number of subjects) the number of available subjects may be increased.

### 2.3.2. Types of speech stimuli

Various types of speech stimuli are used in speech intelligibility tests. Generally, the length of each single stimulus determines what level of analysis (Table 2.1) is addressed by the test method.

The most fitting speech stimuli corresponding to the different levels indicated in Table 2.1 would appear to be sentences, words and phonemes. However, individual phonemes are hard to test without the context of a word or syllable; hence the frequent use of nonsense syllables that was mentioned in the previous section. The individual recognition of phonemes is also difficult to test using meaningful words, since the word context will be of some influence on the probability of correct recognition.

Higher-than-word level effects are expected for most practically feasible cross-language conversations. In principle, sentence intelligibility tests also include effects at lower (word and phoneme) levels, since all sentences are constructed from these smaller units of speech. If only one type of speech stimuli can be chosen, it makes sense to choose sentences. On the other hand, it should be noted that (nonsense) word tests will be more sensitive to effects at lower levels of analysis.

When comparing native and non-native talkers, specific choices must be made before recording any speech stimuli. Speaking rate and speaking style are likely to vary between native and non-native talkers. Non-native talkers usually tend to (consciously or unconsciously) compensate for the effects of their accent on intelligibility by adjusting their speaking rate or speaking style (van Wijngaarden et al., 2001b). This is a legitimate effect, which can also be observed in cross-language conversations in practice—it is in some ways similar to the Lombard-effect, which makes talkers automatically increase their vocal effort in the presence of background noise. One may choose to include this effect in the test, or force native and non-native talkers into similar speaking styles (by giving suitable instructions, monitoring recordings, and pacing their speaking rate).

### 2.3.3. Availability of multiple languages

Multi-lingual intelligibility testing is one step beyond non-native speech intelligibility testing is. Multi-lingual tests can involve either native or non-native subjects, but must also be implemented in multiple languages. Having

a multi-lingual test can be extremely useful for cross-language research. For obtaining L2 results as well as an L1 baseline for a single subject, one needs to have a test that is (at least) bilingual.

Obtaining equivalent implementations of the same test in various languages poses an additional difficulty. True equivalence across languages is hard to reach. Whatever speech stimuli are used, these stimuli must somehow be matched across languages. When working with phoneme tests, the tested phonemes could be balanced to represent the mean frequency of occurrence in the corresponding language. Despite the fact that different phoneme sets must be tested for each language, these are equivalent in the sense that they represent a 'natural' distribution of phonemes for each language.

When the test stimuli are isolated words, then on top of phonetic balancing, the frequency distribution of the test vocabulary (measured frequencies of occurrence in representative texts) should be controlled. Where available, the appropriate information could be taken from (multi-lingual) lexical databases.

When using sentences, the main aspects that should be matched are the complexity of the sentences, and the *domain* from which they are taken. The source of the sentences largely determines the domain (newspaper, radio, everyday conversation, etc.), making this variable relatively easy to control. The complexity can be controlled by adopting certain constraints for the selection of sentences; at least the length (number of syllables) of the sentences, and the length of the individual words, should match pre-defined criteria.

If sentences are properly selected, phonetic balancing becomes of lesser importance. Each sentence consists of a certain mix of phonemes; if each condition is tested with multiple sentences, there is a more or less implicit phonetic balancing for the domain from which the sentences are taken.

An additional complicating factor when designing multi-lingual tests is the fact that the relative importance of different levels of analysis (Table 2.1) may vary between languages. Phoneme identification may be more difficult in some languages than others, simply because the number of phonemes differs (e.g., English vowels versus Spanish vowels). Contextual information that is available in one language, for instance in the form of case and word gender, may be absent in other languages.

A pragmatic approach to the design of multi-lingual tests is to simply try out the implementations in different languages on native subjects. If the native scores are the same across languages, then it seems fair to assume that the method performs equivalently.

## 2.4. METHODS USED IN THIS STUDY

### 2.4.1. Speech Reception Threshold (SRT) method

The SRT method is widely used as a diagnostic tool in the field of audiology (Plomp and Mimpen, 1979), and has been proven useful to evaluate speech intelligibility of talkers, listeners, and communication systems. In order to extend the applicability of the SRT method to cross-language applications, a multi-lingual corpus of test sentences was recorded.

#### *2.4.1.1. Experimental procedure*

The SRT test gives a robust measure for sentence intelligibility in noise, corresponding to the speech-to-noise ratio that gives 50% correct response of short redundant sentences. In the SRT testing procedure, masking noise is added to test sentences in order to obtain speech at a known speech-to-noise ratio. The 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. If every word in the responded sentence is correct, the noise level for the next sentence is increased by 2 dB; after an incorrect response, the noise level is decreased by 2 dB. The first sentence of a list of 13 sentences is repeated until it has been responded to correctly, using 4 dB steps. This is done to quickly converge to the 50% intelligibility threshold. By taking the average speech-to-noise ratio over the last 10 sentences, the 50% sentence intelligibility threshold (SRT) is obtained.

#### *2.4.1.2. Interpretations of SRT results*

The score resulting from an SRT test ('the SRT' for the corresponding condition) is a speech-to-noise ratio (SNR); at this SNR, 50% of the sentences are repeated correctly by the listeners. At better (higher) SNRs, more than 50% will be intelligible, at more adverse (lower) SNRs, less than 50%. A lower SRT means better intelligibility: more noise can be allowed to reach 50% recognition of sentences.

The percentage of correctly recognized sentences is a (psychometric) function of the SNR, often modeled as a cumulative normal distribution. The SRT is the adaptively estimated mean of this distribution, which is the best single parameter to characterize the whole curve. A logical second parameter to estimate would be the variance of the distribution, reflected by the slope of the psychometric curve. To estimate this slope (or even the full psychometric curve), one uses alternative testing paradigms based on the same SRT sentences (descriptions of such paradigms used in this study are given in the corresponding chapters).

#### *2.4.1.3. Creating a multi-lingual test sentence corpus*

The original (Plomp and Mimpen, 1979) SRT sentences describe common, everyday situations in simple wording. These sentences were read in a very clear style by a single female talker. SRT results with the original corpus have been found hard to parallel by newer SRT corpora (e.g., Versfeld et al., 2000). The original SRT corpus is available only in Dutch, although a number of (independent) translations were created. Instead of combining existing SRT implementations in various languages, the original sentences were translated anew, aiming at maximum consistency across languages.

The following constraints were defined for 'translation' of the sentence material:

- Translation need not be literal, but should fit in a similar context as the original
- Sentence length 7–9 syllables
- No words longer than 3 syllables
- No more than one three-syllable word per sentence.
- Sentences of approximately equal redundancy (or predictability, perplexity) as the original sentences

The Dutch sentences were translated directly into German and English, by a native speaker of these languages, and corrected by another native speaker, with a background in phonetics. The multi-lingual SRT corpus meanwhile contains more languages (not used in the experiments described in this thesis), which are mostly translations of the English sentences.

#### *2.4.1.4. Speech recordings*

Traditionally, talkers used in SRT tests for audiological purposes are trained professionals, speaking very clearly. Audiologists are interested in an SRT that offers the sharpest possible criterion for speech hearing acuity; as a result, the choice of talkers and speaking styles may not be very representative of what the average listener is confronted with in the real world.

The talkers for the multi-lingual SRT corpus were not selected according to a strict regime, or following specific criteria. The talkers (of various language backgrounds) were simply verified not to exhibit any speaking disorders, and instructed simply to speak in a clear voice. As a result, most talkers adopt a speaking style that is somewhat conversational, rather than the professional speaking style of a news reader. This approach makes it easier to recruit talkers, and is also more representative of speech heard in real life.

To prevent large differences in speaking rate, the speaking rate is paced by means of a 'progress bar'. Talkers have to pronounce each sentence within a three-second timeframe, which is visually indicated on the computer screen.

#### *2.4.1.5. Software implementation*

A computer program was developed for maintaining multi-lingual databases of recorded SRT sentences and using these in intelligibility tests. This program also features a module for recording new material. In combination with a notebook computer and a high-quality sound card, a small, flexible setup is created, which can be used to record and test non-native talkers and listeners at any location that is sufficiently silent.

### 2.4.2. Semi-open response Consonant-Vowel-Consonant (CVC) method

A type of semi-open-response CVC (consonant-vowel-consonant) intelligibility test was developed for the purpose of testing phoneme intelligibility with non-native subjects. Using this test, recognition of initial consonants and vowels could be scored, and confusion matrices could be composed (cf. Miller and Nicely, 1954). The method is similar to an open-response equally-balanced CVC paradigm (Steeneken, 1992). The main differences are that the final consonant is not tested, and that the subject responds by choosing an alternative from a (nearly) exhaustive list of possible CVC-words, instead of typing the word in response to the stimulus. The advantage of this approach is that extensive training of subjects becomes unnecessary, while the construction of confusion matrices is still possible. Problems that were expected using a 'difficult' open-response paradigm with non-native subjects can be successfully avoided.

Only a Dutch version of this method was implemented, which limits the scope of the test compared to the SRT test.

#### *2.4.2.1. Experimental procedure*

The listener is presented with speech utterances by a talker. Each presentation consists of a single CVC nonsense-word, embedded in a short carrier phrase. A number or alternative CVC-word responses, from which the listener has to choose the one that he or she has heard, are displayed on a computer screen.

Each test run takes 3 to 4 minutes, during which almost all initial consonants and vowels that occur in the Dutch language are tested once. Initial consonants and vowels with a frequency of occurrence (based on the Dutch newspaper "NRC Handelsblad") below 2% were not included in the test, leaving 17 initial consonants and 15 vowels.

When testing an initial consonant, 17 alternatives are displayed, and for a vowel 15 alternatives. Although this is essentially a closed-response approach, it is very unlikely that the listener would like to respond with a

CVC word that is not given as one of the alternatives. The fact that the list of alternatives is nearly exhaustive makes the choice 'semi-open.'

A presentation that is aimed at testing the vowel /Ø:/, for instance, could give the listener the following list of alternatives to choose from: 'jaap', 'jup,' 'jeup,' 'jip,' etc. The only difference among the alternatives is the vowel (rhyme word concept).

Each test run consists of 32 presentations, in random order. CVC-words are formed by combining the 32 different test phonemes (15 vowels, 17 initial consonants), with 32 sets of two non-tested phonemes. These non-tested phonemes, influencing the test through co-articulation effects, were not chosen randomly. Instead, the selection was such, that the spread of these phonemes over a perceptual space (Pols, 1977) was more or less maximized within each single list. From the entire set of possible options for the non-tested phonemes, subsets of phonemes were chosen in such a way, that these subsets were spread out over the entire perceptual space.

To obtain sufficiently reliable results, each phoneme needs to be tested several times in each condition. Between these reproductions, different CVC-words are used.

#### *2.4.2.2. Speech recordings and software implementation*

Speech was recorded for native and non-native talkers of the Dutch language, following a similar computer-paced approach as for the multi-lingual SRT sentences (by showing each utterance on a screen, allowing a fixed time window for the talker to speak, and applying level normalization).

The semi-open CVC tests were controlled by a computer program, using the same hardware as for the SRT test. Confusion matrices and recognition scores were automatically compiled from subject responses.

## 2.5. DISCUSSION AND CONCLUSION

The pragmatic model of cross-language speech communication presented in this chapter was used to select the multi-lingual SRT method as a suitable "general purpose" tool for measuring non-native speech intelligibility. As the experiments described in the following chapters will show, the method is effective in collecting quantitative data for non-native talkers as well as listeners.

The semi-open response CVC test is used as a secondary method, which provides more lower-level details, but may not always be as representative of speech intelligibility in practical scenarios.

# Chapter 3. Quantifying the intelligibility of speech in noise for non-native talkers[2]


## ABSTRACT

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). The 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.


## 3.1. 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, among other things, by differences in articulatory precision (Bradlow et al., 1996), speaking rate (Sommers et al., 1994) and speaking style (Bradlow and Pisoni, 1999;

[2] This chapter is a slightly modified version of a previously published paper: van Wijngaarden, S.J., Steeneken, H.J.M. and Houtgast, T. (**2002**). "Quantifying the intelligibility of speech in noise for non-native talkers," J. Acoust. Soc. Am. **112**, 3004–3013.

Picheny et al., 1985). 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 et al., 2001b).

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. If 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 (see Chapter 2). In principle, segmental as well as supra-segmental influences can be expected. There has traditionally been much attention for 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 overall intelligibility, it is unclear to what degree. The presence of context will enable listeners to correctly interpret many non-authentic 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.

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 described in this chapter are all concerned with speech in the presence of noise. The influence of noise can be seen as representative of many speech degrading conditions.

## 3.2. DEGREE OF PERCEIVED FOREIGN ACCENT

### 3.2.1. Methods

Inexperienced second language (L2) talkers are often recognized for 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 foreign-language environment should be able to assess the strength of their own accent, if only by its apparent effect on native listeners.

#### *3.2.1.1. 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 4 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 3.1).

All talkers were asked to rate their Dutch proficiency on a five-point scale ("bad"–"excellent"), 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 opportunity to revise their self-ratings after the recording session, but none of the talkers chose to do so.

#### *3.2.1.2. 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 from 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 seconds 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 transformed to a *Z*-scale. 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 (Chapter 2 of this thesis). A total of 19 native listeners participated;

10 of these listeners repeated the experiment 3 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 had been tested for normal hearing.

### 3.2.2. Results

In Table 3.1, relevant information regarding the 15 talkers is given, together with proficiency self-ratings and accent ratings from the pairwise comparison experiment.

Table 3.1. 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 with 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 acq. | Expe-rience with Dutch (yrs) | Self-rating for speaking (1–5) | Mean self-rating (1–5) | Overall accent rating (Z-score) |
|---|---|---|---|---|---|---|
| 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 |

As can be seen in Table 3.1, the L2 talkers differ with respect to their experience with the Dutch language. All first started learning Dutch as adults. Hence, the experimental results obtained with these talkers apply to clearly post-lingual 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 3.1, and further established by Fig. 3.1, 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.

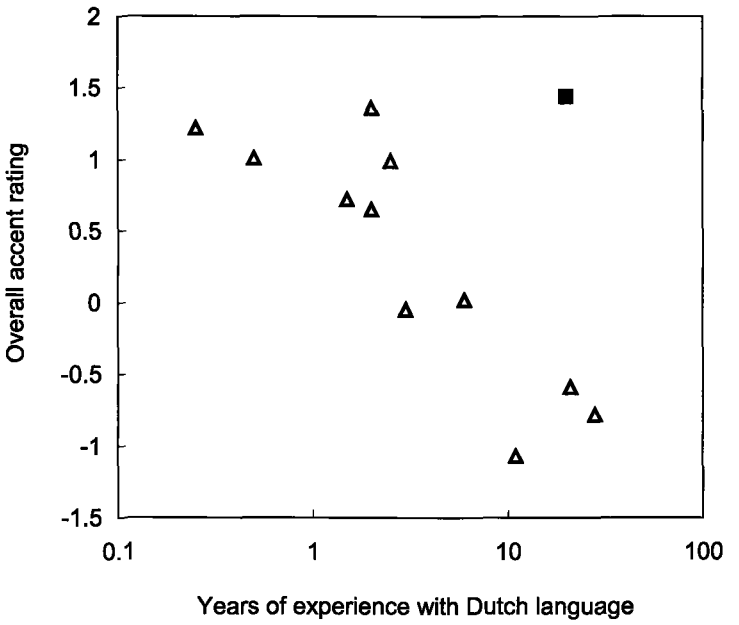


Figure 3.1. Relation between foreign accent ratings and years of experience with the Dutch language, for the 12 L2 talkers. With the exception of talker CF-15 (indicated by a black square) the accent rating correlates well with the logarithm of the number of years of experience ($R^2 = 0.74$, without CF-15).

Please note the logarithmic scale in Fig. 3.1. 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. 3.2.

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

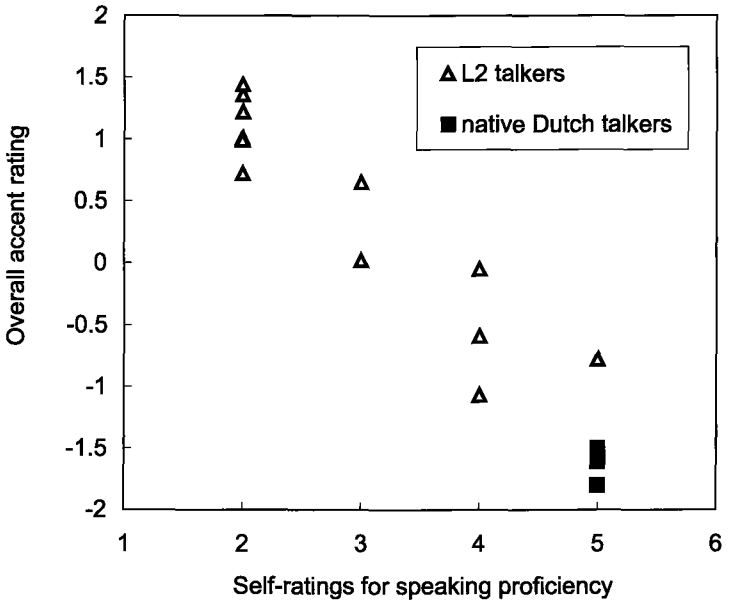


Figure 3.2. Relation between self-ratings for speaking proficiency and foreign accent ratings from pairwise comparisons by native Dutch listeners ($R^2 = 0.91$).

We are mainly interested in the degree of foreign accent for its effect on speech intelligibility. In this light, a limitation of the accent ratings from Figs. 3.1 and 3.2 is that, since the subjects rated accent holistically, various speech characteristics may have contributed to the ratings. For example: a fluent 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 disfluently. 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 10 listeners. They were explicitly asked to compare the pairs of talkers, based on only one of these two dimensions. The same 10 listeners compared all pairs of talkers twice on both dimensions, in consecutive experiments. The relation between the scores from these experiments and the overall accent ratings from the original pairwise comparison experiment is given in Figs. 3.3 and 3.4.

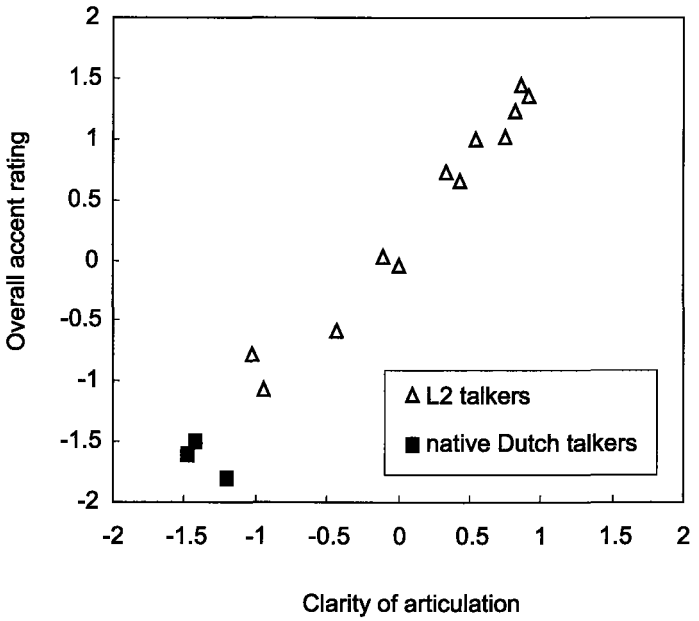


Figure 3.3. Relation between pairwise comparison ratings for 'clarity of articulation' and overall foreign accent ($R^2 = 0.97$).

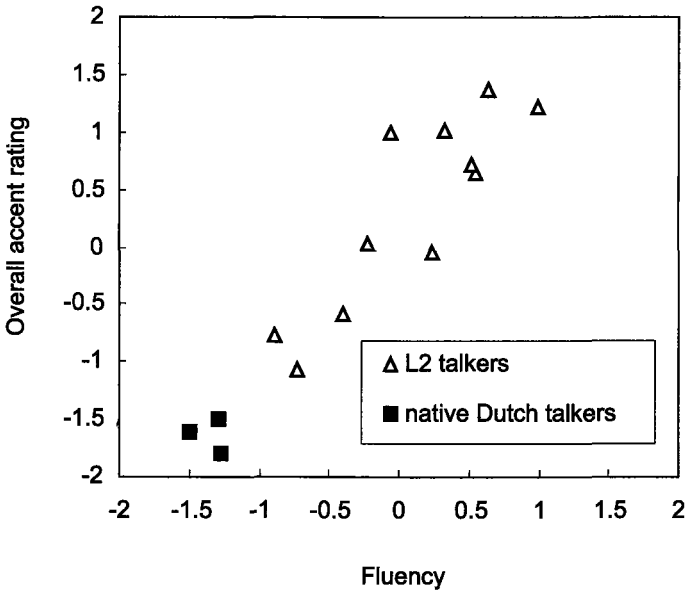


Figure 3.4. Relation between pairwise comparison ratings for 'fluency' and overall foreign accent ($R^2 = 0.89$).

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 predominant factor for the perception of overall accent strength.

For practical reasons, 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 3.2.

Table 3.2. 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* | $r \leq -1$ | $-1 < r \leq 0$ | $0 < r \leq 1$ | $r > 1$ |
| 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 |

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

### 3.3.1. Methods

We expect non-native speech production to be influenced by factors at segmental and supra-segmental level. If 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). A suitable method for investigating speech intelligibility at the phoneme level is the semi-open response Consonant-Vowel-Consonant test (see Chapter 2).

#### *3.3.1.1. 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 SRT listeners. Of these listeners, 16 also participated as listeners in the CVC experiment

Because of the time-consuming nature of the CVC 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.

#### *3.3.1.2. 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 (examples of psychometric functions can be seen by skipping ahead to Fig. 3.6).

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 85–100% 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.

### 3.3.2. Results and discussion

#### *3.3.2.1. SRT scores for non-native talkers*

Speech reception thresholds for each of the twelve L2 talkers, as measured with 20 native listeners, were all equal to or higher than those for the 3 native talkers. This means that the intelligibility of the L2 talkers is, as expected, equal to or lower than that of native speakers of the Dutch language. The mean SRT score for each talker is given in Table 3.3. Note that the standard error reported in this table, which is an indication of statistical accuracy of the estimated SRT, is different from the standard deviation (slope) of the psychometric function mentioned in 3.3.1.2.

Table 3.3. 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 |

The relation between perceived foreign accent and speech intelligibility is shown in Fig. 3.5.

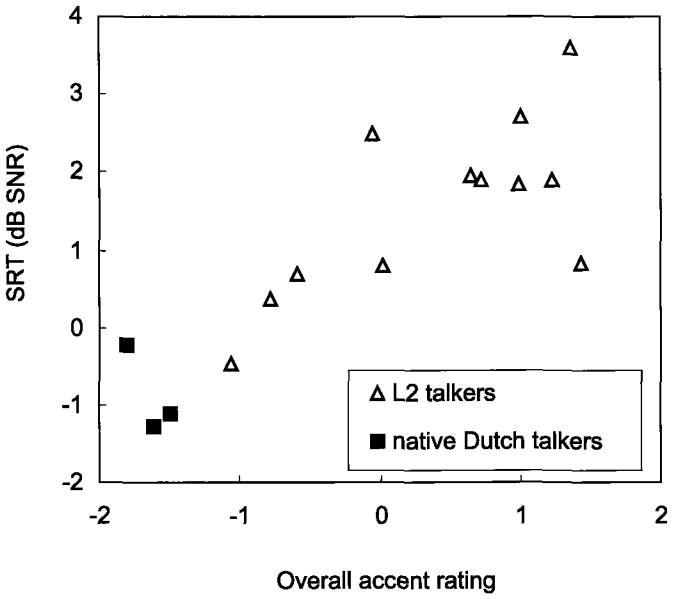


Figure 3.5. Relation between foreign accent ratings and SRT scores for speech intelligibility. Accent strength is significantly correlated with speech intelligibility ($R^2 = 0.70$).

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 observed 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. 3.5) as well as self-ratings hold a predictive value for speech intelligibility.

When comparing Table 3.3 and Fig. 3.5 to similar data for non-native listeners instead of talkers (Chapter 4 of this thesis), it appears that the effect of non-native speech production on intelligibility tends to be smaller than that of non-native perception.

#### *3.3.2.2. Slope of the psychometric function for sentence reception*

Since perceived accent strength and intelligibility correlate well, it can be assumed that the division into accent strength categories given in Table 3.2 holds as a division in categories for intelligibility effects. Therefore, the slope of the psychometric function for sentence recognition was not measured for all talkers, but only for one talker from each category. An exception was made for native talkers; all three of these were included, in order to be able to get an impression of regular (native) inter-speaker variability. The means of the psychometric functions and the slopes around the 50%-points, measured using the procedure as described in 3.3.1.2., are given in Table 3.4.

Table 3.4. Mean (SRT) and slope op the psychometric function for sentence recognition in noise. Means and standard errors across 5 listeners are given. The slope for the cat. III and IV talkers differs significantly from the slopes for the cat. I talkers, the slope for the cat. II talker does not.

| 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 |

Please note that the 50%-point of the psychometric function as reported in Table 3.4 is essentially the same measure as the SRT reported in Table 3.3, but determined with another paradigm. The correspondence between these values for the same talkers is good.

Table 3.4 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. 3.6, which shows the full psychometric functions according to the data in Table 3.4, assuming that these follow a cumulative normal distribution.

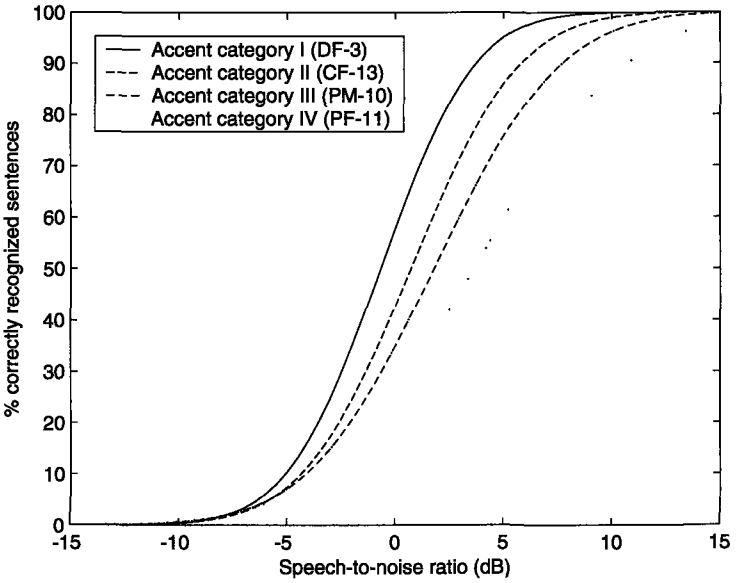


Figure 3.6. Average psychometric functions for the recognition of sentences by four talkers, differing in accent strength.

Figure 3.6 clearly shows that the 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.

#### *3.3.2.3. Phoneme recognition*

So far, all presented speech intelligibility data have been 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 relationship 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 effect with the effect of talker proficiency, only Polish talkers are compared to a (baseline) Dutch talker (see Figs. 3.7 and 3.8).

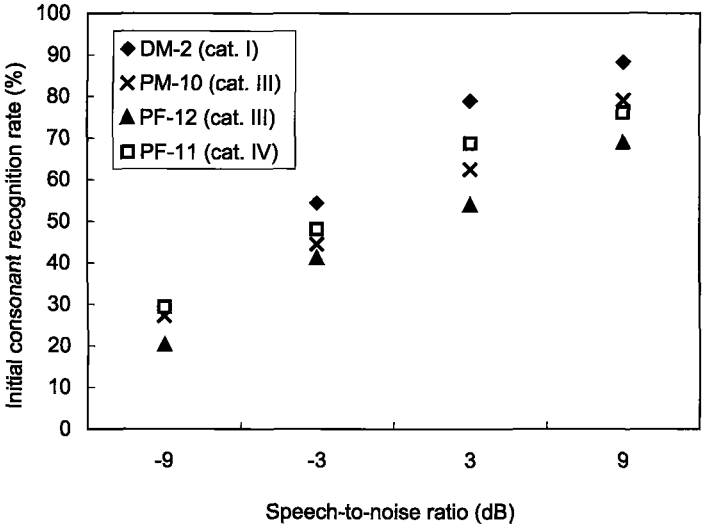


Figure 3.7. Percentage of correctly recognized initial consonants in CVC words for three Polish and one Dutch talker speaking Dutch, as a function of speech to noise ratio (mean values across 16 native listeners; standard errors are in the range of 2–4.5 percentage point).

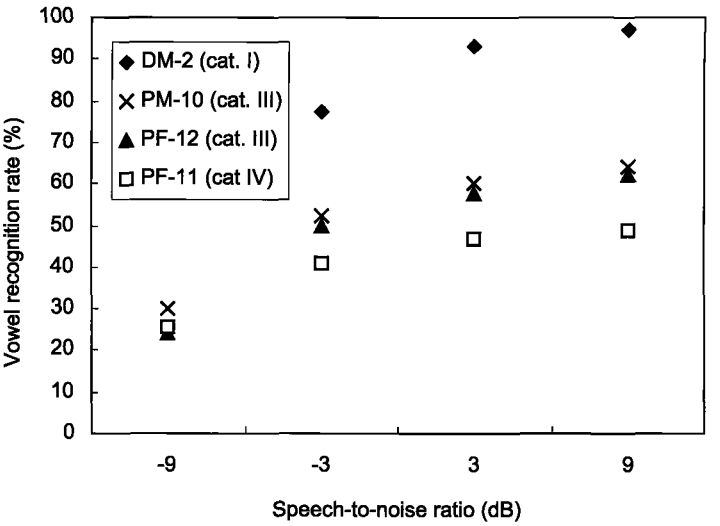


Figure 3.8. Percentage of correctly recognized vowels in CVC words for three Polish and one Dutch talker speaking Dutch, as a function of speech to noise ratio (mean values across 16 native listeners; standard errors are in the range of 2.2–5.3 percentage point).

There is a clear (and statistically significant) overall effect of foreign accent on initial consonant recognition (Fig. 3.7), 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 has not yet been reached.

The recognition of individual vowels (Fig. 3.8) appears to be explicable 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 these talkers are from the same language background, one might expect that they all have difficulties pronouncing the same vowels. The Polish vowel system has 8 vowels, of which 6 (/Iieaou/) also occur in Dutch (e.g., Martens and Morciniec, 1977), and are included in the CVC test. Individual realizations of these vowels differ 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* in both languages, but are always at least *similar*. The other 9 vowels included in the Dutch CVC test (including three diphtongs) do no occur in Polish at all.

To see if the patterns of vowel confusion are consistent across talkers, the percentage of correct recognition was 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.

Table 3.5. 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 | – |

As Table 3.5 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 or SLM (Flege, 1992; Flege, 1995) predicts that late L2 learners, such as the Polish talkers in our experiments, are able to master 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 9 new versus the 6 similar vowels. In Fig. 3.9, the scores for 'new' and 'similar' vowels are given for the different talkers.

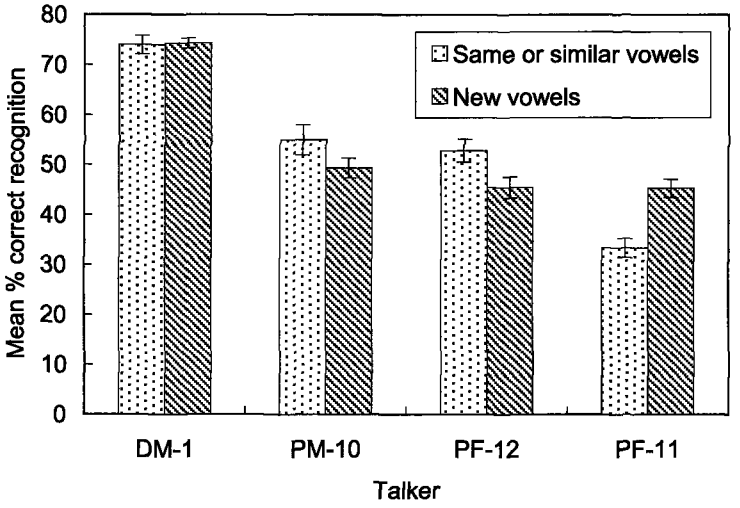


Figure 3.9. Percentage of correctly recognized vowels for two sets of vowels: Dutch vowels that are the same (or similar) in Polish, and Dutch vowels that are new to Polish learners of the Dutch language. The error bars indicate the standard error ($N$ = 16; mean percentages taken per listener).

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).

#### *3.3.2.4. Relation between phoneme and sentence intelligibility*

The overall recognition of sentences (Fig. 3.6), although fundamentally based on phoneme recognition, follows a somewhat different pattern than the recognition of individual phonemes (Figs. 3.7 and 3.8). The difference that is perhaps noted first, is that ceiling effects as observed for vowel recognition appear absent from sentence recognition results[3]. 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 usually 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 of the intelligibility of entire messages. Ceiling effects do, in this case, occur for vowels (Fig. 3.8), 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 of the intelligibility of messages. This is further illustrated by Fig. 3.10, which combines data from Figs. 3.6 and 3.8.

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

[3] One could argue that the psychometric functions of Fig. 3.6 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.

phoneme-based measures that are known to predict sentence intelligibility in native speech, require validation before applying those measures to non-native speech.

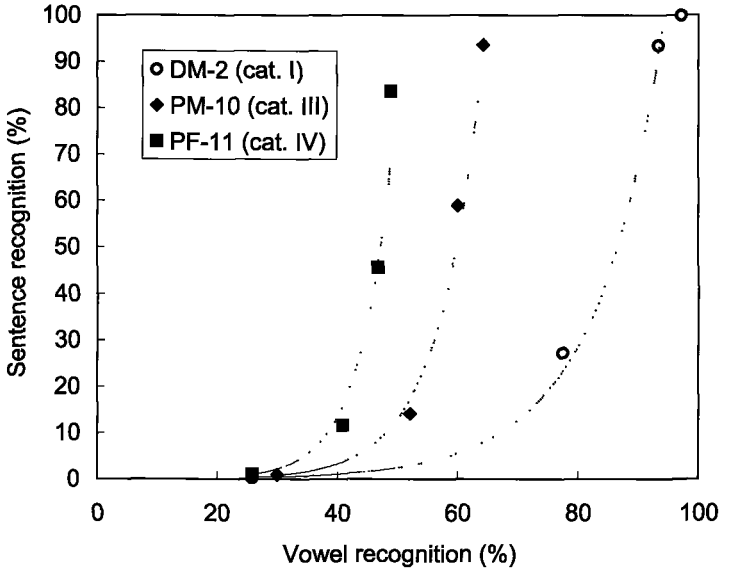


Figure 3.10. Sentence recognition as a function of vowel recognition for three talkers (one native, two non-native), at 4 different speech-to-noise ratios (–9, (–3, 3 and 9 dB). To guide the eye, an exponential curve has been fitted to the data of each talker.

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

### 3.4.1. 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 chapter, 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 3.3) were used. The SRT sentence recordings had been paced by means of a visual time indicator, allowing the talkers up to 2.6 seconds 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 second '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 relationship 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 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.

### 3.4.2. 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 (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. (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 what 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 3.6.

Table 3.6. Percentage of vowels (10 vowels, 3 realizations each) correctly classified according to the vowel regions by Pols, Tromp and Plomp (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 |

The scores are higher for the two male native talkers than for the non-native talkers. The mean percentages of correct classification per vowel, for all of the talkers in Table 3.6, were subjected to a 2-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.

## 3.5. DISCUSSION AND CONCLUSIONS

Foreign-accented speech tends to be less intelligible than native speech. The results presented in this chapter 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 chapter 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.

# Chapter 4. Quantifying the intelligibility of speech in noise for non-native listeners[4]


## ABSTRACT

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


## 4.1. INTRODUCTION

Most people know from personal experience that non-native speech communication is generally less effective than purely native speech communication. This is readily verified by listening to foreign-accented

[4] This chapter is a slightly modified version of a previously published paper: 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.

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

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

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

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

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

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

Experiments concerning non-native speech intelligibility in noise will be described in Section 4.2: Speech Reception Threshold (SRT) results are presented, which will allow a broad quantitative comparison between native and non-native speech intelligibility in noise. In Section 4.3, this comparison will be refined by looking at the slope of the psychometric function in a sentence recognition task. Section 4.4 will describe experiments exploring the relation between non-native sentence recognition and redundancy-related measures.

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

### 4.2.1. Method

An interesting topic in relation to non-native speech perception is the use of word context. This means that speech intelligibility for non-native listeners is best measured using longer phrases (sentences). For measuring sentence intelligibility under the influence of noise, several proven methods are available, among which the Speech Reception Threshold (SRT; Plomp & Mimpen, 1979). The SRT method was used for all intelligibility experiments described in this chapter

#### *4.2.1.1. Subjects*

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

Two main groups of subjects were recruited for this experiment. Group I was recruited following fairly strict guidelines. The recruiter used a

'checklist' to make sure that only subjects were accepted that matched a set of pre-defined criteria. Group I consisted of 9 tri-lingual Dutch university students of various disciplines (not including languages or phonetics), aged 18–24, who considered English their second language and German their third language. All had first learned both English and German, written and orally, during secondary education (Dutch high school), all starting with English at age 12 or 13, and with German at age 13 or 14. For each individual subject, the self-reported overall proficiency (rated on a 5-point scale) was higher for English (mean rating 3.7) than for German (mean rating 2.9). All individual subjects had a much more frequent use of English than of German: all reported daily use of English (reading and/or listening), while use of the German language was typically weekly to monthly.

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

In addition to the main subject groups I and II, two control groups were recruited: 3 native German and 3 American subjects. These control groups were used to verify that the implementation of the SRT test (sentence material and talkers) was equivalent across languages.

#### *4.2.1.2. Stimuli*

Following the procedures described in Chapter 2, a set of 130 Dutch SRT sentences (10 lists of 13 sentences) were 'translated' into German and English. The sentences were recorded as spoken by native talkers of Dutch, German and American English. Additionally, the same Dutch talkers (who were tri-lingual) also recorded English and German sentences. Recordings were made for a total of 9 talkers: 3 for each native language (2 male, 1 female); because of the fact that the Dutch talkers recorded three sets of sentences (in Dutch, German and English), a total of 15 sets of recorded sentences was collected.

Talkers did not demonstrate any speaking disorders, and were informally estimated to have more or less average clarity of articulation.

Influences of regional accents (deviations from the preferred pronunciation in the respective languages), when noticeable at all, were minor.

### 4.2.2. Results

#### *4.2.2.1. Fully native baseline SRT-scores*

Conclusions regarding the effects of non-nativeness can only be drawn, if the SRT implementation that is used is also independent of language. In other words: we need to make sure that the precautions taken in the 'translation' of the test sentences were effective in making the German and English test equal to the original Dutch test. This was verified by conducting 'fully native' SRT tests in all three languages (3 talkers per language; 3 English listeners, 3 German listeners and 20 Dutch listeners).

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

Compared to SRT-results found with thoroughly optimized SRT databases, a mean SRT of –1 dB may seem high. For an non-optimized SRT test in Dutch (but *with* specifically selected talkers, which is not the case in the multi-lingual SRT test), Versfeld et al. report a mean SRT of –1.8 dB (Versfeld et al., 2000). The difference can most likely be attributed to the concessions done to keep the recording procedure practical, a more informal speaking style (see Chapter 6), and the absence of a strict talker selection regime (see Chapter 2).

#### *4.2.2.2. SRT-scores of group I.*

Group I, the homogeneous group of 9 tri-lingual Dutch subjects, participated in an SRT experiment in which subjects were presented with Dutch (D), German (G) and English (E) speech. In addition to the SRT sentences by (native) G and E talkers (G>(G)>D and E>(E)>D), they were also presented with speech by the three Dutch talkers in German and English (D>(G)>D and D>(E)>D). In this latter case, the overall intelligibility will not only be affected by non-native speech perception, but also by non-native speech production. The results from this experiment, separated by individual talker, are given in Fig. 4.1.

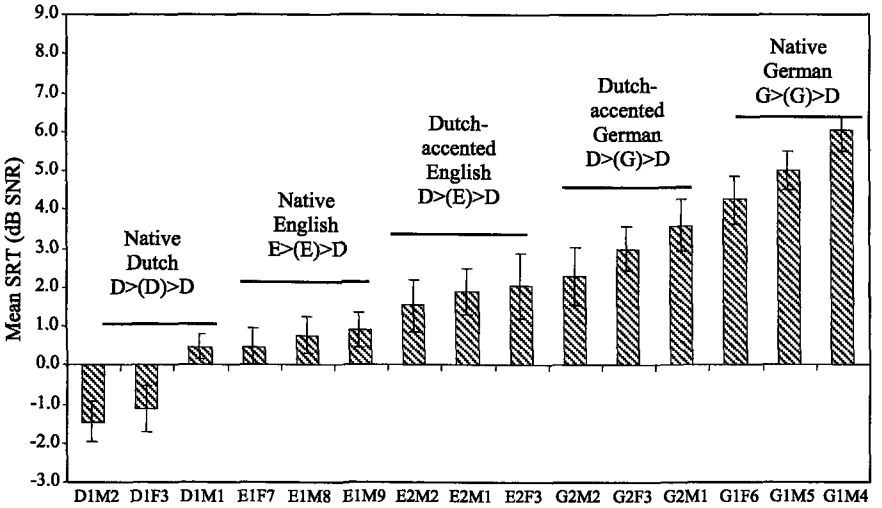


Figure 4.1. Mean SRT results of subject group I per individual talker ($N$ = 9). All listeners were Dutch students, speaking English as second language and German as third language. Speech material was in Dutch (D), English (E) and German (G). Non-native talkers (D>(E)>D and D>(G)>D) were all Dutch. The talkers are labeled according to language group (e.g., D1 for native German, E2 for non-native English), gender (M for male, F for female), and a unique number (1-9) to identify each individual person acting as a talker. The error bars represent the standard error.

The talkers in Fig. 4.1 are grouped by language, and rank-ordered according to mean SRT for all 9 listeners. The effect of non-native perception for English (difference between D>(D)>D and E>(E)>D scores) is relatively small; the mean difference in SRT is 1.4 dB. The mean difference between D>(D)>D and G>(G)>D is much larger: 5.8 dB. Different deficits for English and German were to be expected; differences in proficiency and intensity of use have a clear effect on intelligibility. Compared to earlier results from similar studies in other languages (Buus et al., 1986; Mayo et al., 1997), the G>(G)>D deficit matches expectations, but the E>(E)>D deficit is smaller than expected for late bilinguals. The frequent 'early' exposure of young Dutch people to English speech on television may be part of the explanation.

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

#### *4.2.2.3. SRT-scores of groups I and II together (group I+II)*

The same SRT-conditions presented to group I were also tested with group II[5]. By combining the data of groups I and II, analyses based on a larger group of 20 subjects (which we will call 'group I+II') may be carried out, which will be more diverse in terms of their proficiency, at least in German. This allows us to study the effect of proficiency and experience on speech intelligibility.

In Figs. 4.2 and 4.3, combined SRT-results for group I+II are given. Scores for the 20 subjects were divided into 4 subgroups of 5 subjects, according to the self-reported proficiency of the subjects. The leftmost subgroup in each figure is the subgroup with the lowest self-reported proficiency, the rightmost is the one with the highest proficiency. Although Fig. 4.2 (English) and Fig. 4.3 (German) are based on scores of the same 20 subjects, the division into subgroups is different. The division enables investigation of the effect of proficiency on intelligibility. This is not easily done on the basis of individual proficiency ratings, since these tend to be fairly unreliable.

The results of Figs. 4.2 and 4.3 are *not* simply mean SRT scores on the German and English sentences, but rather the *difference* of these scores with the scores on the Dutch sentences (difference with D>(D)>D). This difference is a direct measure of the effect of non-nativeness on speech intelligibility. By taking this difference, a correction is also applied for small differences in (native) Dutch SRT scores between the sub-groups.

Figure 4.2 shows no significant effects of self-reported proficiency. All subjects (also from group II) showed a good command of the English language.

Whereas Fig. 4.2 does not show any systematic relation between intelligibility and self-reported proficiency, Fig. 4.3 demonstrates that such a relation can exist. For authentic, unaccented German speech (G>(G)>D), the intelligibility is higher (the effect of non-nativeness smaller) to the subgroups with higher proficiency ratings. The most proficient subgroup, for example, shows a significantly smaller effect ($p<0.05$) than all of the other three G>(G)>D subgroups. With the exception of the differences between neighboring subgroups, all other G>(G)>D differences in Fig. 4.3 are also statistically significant ($p < 0.05$; *t*-tests used to compare the means between subgroups).

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

---

[5] Compared to group I, group II was (as intended) a less homogeneous group of listeners. Because of the considerable variation in profiency in this group, mean results across group II are not very informative, and are not reported here.

According to Fig. 4.2, E>(E)>D (authentic American English pronunciation) tends to lead to somewhat higher intelligibility than (accented) English speech by Dutch talkers (D>(E)>D). This same effect was observed in Fig. 4.1, and appears to be relatively independent of (small) differences in proficiency.

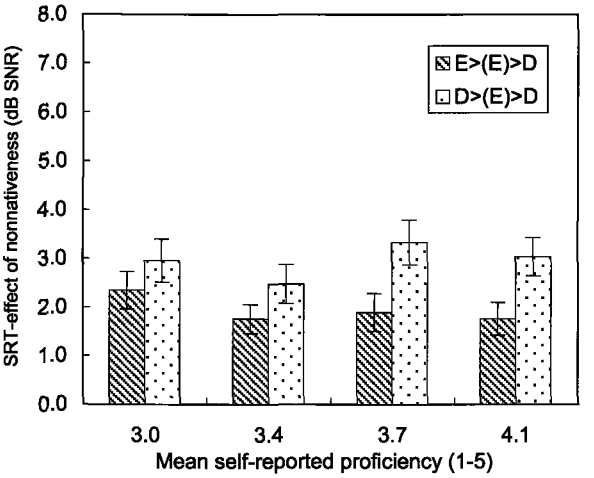


Figure 4.2. The effect of non-nativeness (difference between native and non-native SRT), for subgroups of 5 subjects differing in self-reported proficiency. The non-native language is English. The error bars indicate the standard error (5 subjects, 3 speakers; $N = 15$).

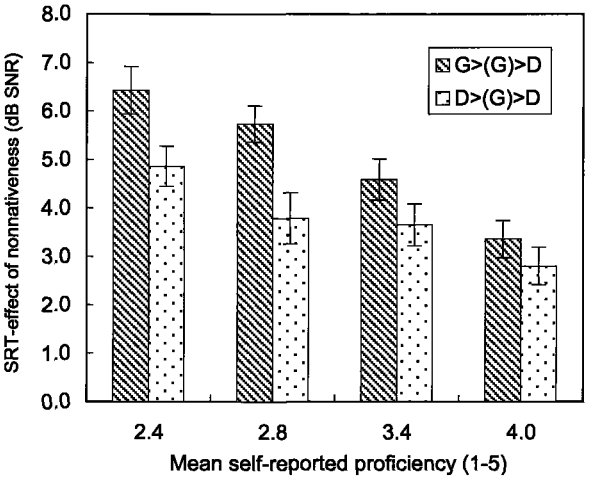


Figure 4.3. The effect of non-nativeness (difference between native and non-native SRT), for subgroups of 5 subjects differing in self-reported proficiency. The non-native language is German. The error bars indicate the standard error (5 subjects, 3 speakers; $N = 15$).

Figure 4.3 shows, much the same as Fig. 4.1, a difference between G>(G)>D and D>(G)>D that is contrary to the difference between E>(E)>D and D>(E)>D. The difference between G>(G)>D and D>(G)>D appears to decrease with proficiency. The two subgroups with the lower self-reported proficiency differ significantly between G>(G)>D and D>(G)>D, the differences are not significant for the other two (more proficient) subgroups.

It is clear that even subjects that give themselves high ratings for German proficiency, have more problems understanding spoken German than the average subject has understanding spoken English. This is observed by comparing the effect of non-nativeness of the *most* proficient (rightmost) subgroup in Fig. 4.3 (German) to the *least* proficient (leftmost) subgroup in Fig. 4.2 (English): the performance in English appears to be still better than in German, although it is difficult to establish clear statistical proof for this.

Please note that the mean proficiency ratings for the sub-groups are only used as relative rankings of proficiency to obtain a division into subgroups. These ratings hold no absolute value; the ratings for English may, for instance, not be directly compared to the ratings for German. The reason for this is that the subjects tend to rate themselves in relation to the performance of their peer group. A more objective measure of proficiency is needed to understand how the results reported in Fig. 4.3 are related to the results in Fig. 4.2 (in other words: how the differences in effects between English and German are explained in terms of differences in proficiency). This will be further explored in Section 4.4.

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

### 4.3.1. Methods

The SRT results given in Section 4.2 characterize the psychometric function of sentence intelligibility by a single value: the SNR for which 50% sentence recognition occurs. However, much speech communication in real life takes place at speech-to-noise ratios corresponding to other levels of sentence intelligibility than 50%. We would therefore like to know the full psychometric function, so that we can predict the SNR necessary to meet *any* intelligibility criterion. This is especially relevant since the slope of the psychometric function is known to differ between native and non-native listeners (e.g., Mayo et al., 1997).

The straightforward way of obtaining a full psychometric function is by sampling the curve at a fixed set of speech-to-noise ratios. This can be a rather laborious process. There is a theoretical possibility to extract additional information about the psychometric function from standard SRT measurements (Plomp and Mimpen, 1979). Unfortunately, the SRT

experiments underlying Figs. 4.1 and 4.2 do not include enough individual subject responses at various SNR values to allow an accurate estimate of the steepness of the psychometric function.

A compromise between sampling the entire psychometric function and estimation of the steepness from standard SRT tests was chosen: first the standard SRT was measured, then the percentage of correctly responded sentences was measured directly at 4 speech-to-noise ratios around the SRT. Next, the psychometric function was fit through these points.

The estimated slopes of the psychometric function will be compared across languages. In a fully native setting (talker *and* listener), the SRT in Dutch, English and German was found to be equal (see Section 4.2.2.1), leading to the conclusion that SRT results can be compared across languages in a straightforward way. For the slope of the psychometric function, this firm baseline was not established, but there are no reasons to expect considerable differences.

#### *4.3.1.1. Subjects, stimuli and conditions*

A new group of 15 tri-lingual subjects was recruited, matching subject group I (9 subjects) on all relevant parameters. Since SRT subjects must be unacquainted with the sentence material, and the available material was limited to 10 lists per language, the subjects from experiment I could not participate in this experiment. For the same reason (also given the fact that each individual psychometric function measurement requires the use of five SRT lists), the conditions tested in this experiment do not include all talkers from experiment I.

The three (baseline) Dutch talkers were included, as well as talker E1M8 (see Fig. 4.1) to represent the English talkers and talker G1M5 to represent the German talkers. Dutch talker number 3 was also included as an L2 talker of German (labeled G2F3 in Fig. 4.1) and English (E2F3). Material of each talker was presented to five subjects out of the group of 15.

#### *4.3.1.2. Procedure*

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

Following this procedure, 5 points of the psychometric function were obtained (including the SRT at 50%) per subject per condition. A cumulative normal distribution was fit through these points using a non-linear least-squares approach (Gauss-Newton method). Hence, the model assumed for the psychometric function was a cumulative normal distribution. Effectively,

two parameters of the distribution were fit: the mean and the standard deviation. The mean of the distribution corresponds to the SRT, while the steepness of the psychometric function at 50% intelligibility is directly related to the standard deviation (Versfeld et al., 2000). The steeper the psychometric function, the stronger the effect of a difference in speech-to-noise ratio on speech intelligibility.

### 4.3.2. Results

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

The estimated slopes of the psychometric function around 50% intelligibility are given in Fig. 4.4.

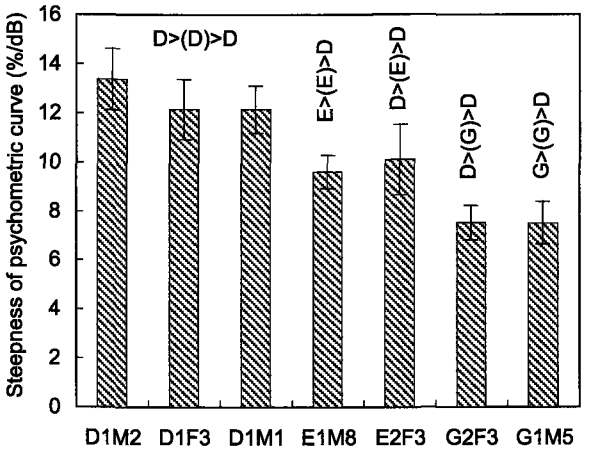


Figure 4.4. Estimates of the steepness (slope at the 50% point) of the psychometric function for 7 individual talkers. Error bars indicate the standard error of the estimates (5 subjects; $N = 5$). The non-native communication scenario corresponding with each talker (e.g., D>(D)>D) is also indicated.

Even at first sight, the steepness of the psychometric function clearly has an inverse relation with the SNR at the 50% point: talkers with higher values of the SRT (see Fig. 4.1) have lower steepnesses, while language appears to be the explaining variable.

The statistical significance of the differences in Fig. 4.4 was investigated by means of a Newman-Keuls test, after finding a significant effect in a one-way ANOVA. None of the differences between talkers speaking *the same language* was significant. The difference between E2F3 and D1M1 is also not significant. All other differences in Fig. 4.4 are statistically significant ($p < 0.05$).

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

In terms of the 50% point of the psychometric function, non-authentic pronunciation was found to be beneficial to Dutch listeners of German, but not of English (Fig. 4.1). Similar effects are *not* found on the slopes of the psychometric function.

## 4.4. RELATION BETWEEN ACOUSTIC AND NON-ACOUSTIC FACTORS

### 4.4.1. The influence of context effects on SRT tests

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

A straightforward measure of linguistic redundancy is obtained through the Letter Guessing Procedure (Shannon and Weaver, 1949), which uses orthographic presentations of sentences to obtain an estimate of linguistic entropy. Other suitable measures, such as the *j*- and *k*-factor (Boothroyd and Nittrouer, 1988) and the c-parameters in the context model by Bronkhorst et al. (Bronkhorst et al., 1993; Bronkhorst et al., 2002), require more complicated and cumbersome experiments.

### 4.4.2. Linguistic Entropy (Letter Guessing Procedure)

The Letter Guessing Procedure (LGP) yields a measure of linguistic entropy (LE); this may be seen as the inverse of the effective redundancy through linguistic factors in the speech material. This measure has been used as a measure of individual subjects' linguistic skills (van Rooij, 1991). Linguistic entropy has been implicated as a predictor of linguistic factors on speech intelligibility (Müsch and Buus, 2001a; van Rooij, 1991).

Since the procedure is based on orthographic presentations of test sentences, what it measures is by definition non-acoustic. Although it is possible to derive redundancy-related measures from spoken language tests, the LGP has some advantages. Because of the orthographic presentation, there are no individual talker effects, and the influence of speech acoustics is eliminated. Furthermore, redundancy at the sub-word level is included, since

individual letters have to be guessed. For practical reasons, this is hard to achieve in any spoken language test, especially with non-native subjects. The orthographic approach also has clear disadvantages. Some factors that are irrelevant for spoken language intelligibility, such as spelling, are included. Also, some very relevant factors, such as phonological transition rules, are not incorporated in the test. However, it is fair to assume that linguistic entropy according to our definition may serve as an indicator of linguistic factors involved in speech recognition.

#### *4.4.2.1. Subjects and stimuli*

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

#### *4.4.2.2. Procedure*

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

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

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

Rather than working directly with the percentage of correctly responded letters, the LGP scores are expressed in terms of linguistic entropy. Entropy, in the context of information theory, is expressed in 'bits'. The linguistic entropy $L$ is related to the fraction of correctly responded letters $c$ according to[6]:

$$L = -\log_2(c) \tag{4.1}$$

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

Assuming a 27-letter alphabet (including space), the linguistic entropy associated with pure guessing of a single letter is, according to Eq. 4.1, 4.75 bits. This is the upper limit to $L$. If all letters are immediately guessed correctly, then $L = 0$: the material is perfectly redundant.

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

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

$$L_n - L_{n-1} = \frac{\alpha}{n} \tag{4.2}$$

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

$$L_n = \beta + \alpha \ln n \tag{4.3}$$

Here the constant $\beta$ may be interpreted as the LE of a single word without sentence context; the constant $\alpha$ quantifies the effect of word position within a sentence on word LE. An exception is made for the first word ($n = 1$), for which Eq. 4.3 is not necessarily expected to hold. Within a set of sentences of a specific structure that is known to the subjects (such as SRT sentences), the predictability of the first word may be much higher than expected from Eq. 4.3.

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

### 4.4.3. Results

#### *4.4.3.1. Relation between LE and SRT for native speech*

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

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

Two new sets of Dutch SRT sentences were constructed, one consisting of proverbs (higher than normal redundancy), the other consisting of semantically unpredictable sentences (SUS), which have lower than normal redundancy (Benoît et al., 1996). LGP and SRT experiments were carried out with 5 native Dutch students, matching subject group II. Individual LE and SRT results are given in Fig. 4.5.

Figure 4.5 shows some residual between-subject variance on the SRT scores, not explained by linguistic entropy. Still, the relation between SRT and LE across sentence types is clear. This means that differences in SRT can be predicted, to a certain degree, from linguistic entropy estimates. The mean increase in SRT as a function of LE is 10 dB/bit between the proverbs and the standard sentences. Between the standard sentences and the semantically unpredictable sentences, this slope is also 10 dB/bit.

The linguistic entropy of the three types of sentences was also calculated for individual words as a function of word position; results of this calculation are given in Fig. 4.6. The very first word of each sentence was not included in this analysis; its baseline-predictability is much higher than all the other words, since it is nearly always an article.

Figure 4.6 shows that LE decreases monotonically with word position, as expected. The estimated values of parameters $\alpha$ and $\beta$ from Eq. 4.3 are given in Table 4.1.

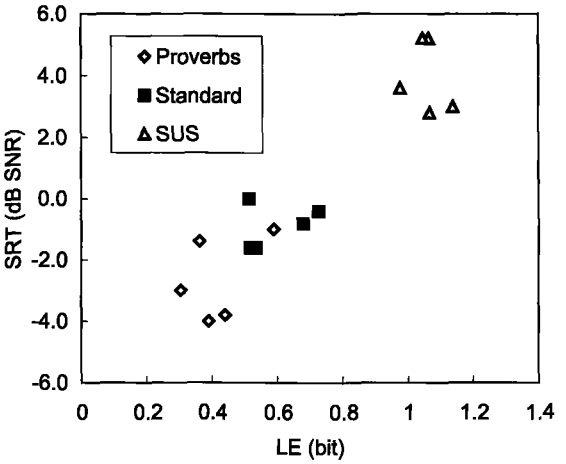


Figure 4.5. Relation between SRT and LE, for 5 individual subjects and three types of SRT sentences. Results are mean values ($N$ = 2 for SRT, $N$ = 13 for LE). Speech material by the same talker was used for all SRT tests.

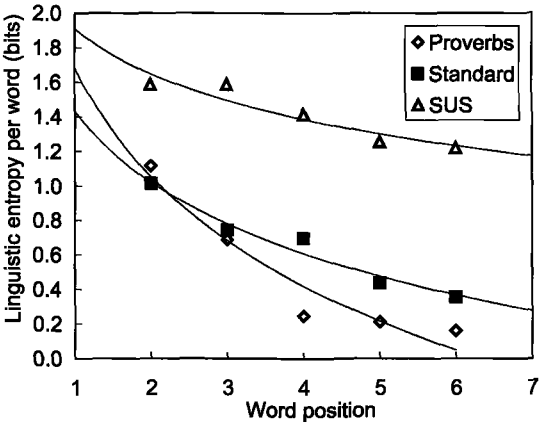


Figure 4.6. Word-LE as a function of word position within sentences, for word positions $2 \leq n \leq 6$. The dashed lines are least-squares fits of equation (3) to the data for the three different kinds of sentences. Data points are based on 5 subjects (each 13 sentences) for Proverbs and Semantically Unpredictable Sentences, and on 9 subjects (each 39 sentences) for the standard SRT sentences.

Table 4.1. Estimated LE parameters from native LGP experiments for three types of sentences.

| Sentence type | Slope $\alpha$ | Offset $\beta$ | $R^2$ (explained variance) |
|---|---|---|---|
| Proverbs | –0.91 | 1.69 | 0.93 |
| Standard SRT | –0.58 | 1.41 | 0.97 |
| SUS | –0.38 | 1.91 | 0.88 |

If it is true that the three types of sentences differ primarily in semantic constraints, then we expect similar values of $\beta$, but different values for $\alpha$. The differences in $\alpha$ are, as expected, statistically significant. However, the differences in $\beta$ are also significant. This may indicate that, between the different sentence types, factors other than semantics were also different, such as word choice (mean frequency of occurrence in natural language, mean familiarity). It could also indicate that the assumption expressed by Eq. 4.2 is not completely justified for words at the beginning of sentences.

### *4.4.3.2. Non-native LE results*

With non-native listeners, linguistic entropy was not varied by manipulating the speech material; instead, it varied according to subjects' individual command of their second or third language. The LGP results of subject group I are presented in Fig. 4.7. Please note that the error bars in Figure 4.7 indicate the standard deviation rather than the standard error, because of the large number of observations per language.

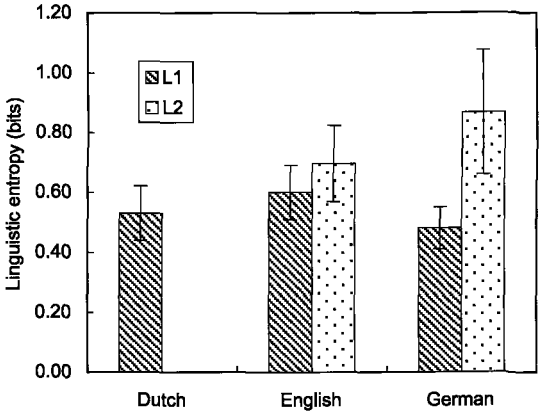


Figure 4.7. Mean LGP results of L2 Dutch subjects (group I) and L1 German and American subjects. All L2 results and L1 Dutch results are based on 9 listeners (39 sentences per listener, $N$ = 351); the L1 German and English results are based on 3 subjects (39 sentences, $N$ = 117). The error bars indicate the standard deviation.

All differences in Fig. 4.7 are highly significant ($p < 0.001$). Unfortunately, and unlike the SRT results, the native (L1) LE scores are *also* significantly different between languages for L1 subjects. Hence, the LGP test is language dependent, and linguistic entropy estimates may not be compared across languages without applying corrections for differences in the LGP test.

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

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

#### *4.4.3.3. Relation between LE and SRT for non-native listeners*

The effects of non-nativeness on LE appear to follow the same patterns as the SRT effects. This suggests that overall intelligibility is largely determined by linguistic factors. Figure 4.8 shows the correlation between normalized LE and SRT, for the individual subjects of group I+II (20 subjects) in all tested languages.

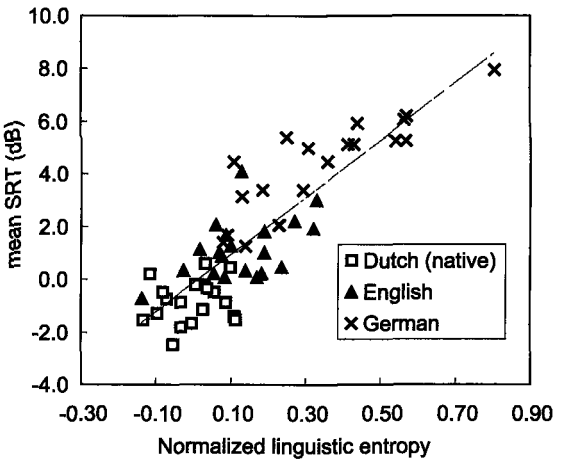


Figure 4.8. Correlation between normalized LE and mean SRT (3 talkers), for native Dutch and non-native English and German (20 subjects). All talkers were native. The dashed line is obtained through linear regression ($R^2 = 0.74$; slope 10.8 dB/bit, intercept –0.15 dB).

The value of the squared correlation coefficient ($R^2 = 0.74$) indicates that roughly 74% of the total variance in SRT scores in Fig. 4.8 may be explained using normalized linguistic entropy. This indicates that LE scores from letter guessing experiments can be used to obtain a fair prediction of corresponding SRT values.

More may perhaps still be learned from mean word-LE as a function of word position, and by estimating the parameters $\alpha$ and $\beta$ of Eq. 4.3. For the subjects of group I, we may verify the effect of the known difference in proficiency between (native) Dutch, English and German (Fig. 4.9 and Table 4.2).

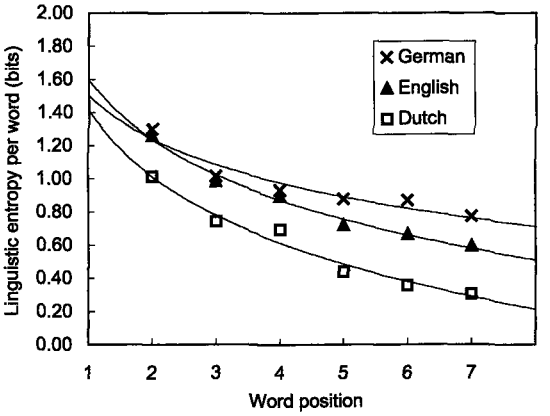


Figure 4.9. Group I word-LE as a function of word position within sentences, for word positions $2 \leq n \leq 7$. The dashed lines are least-squares fits of Eq. 4.3 to the data for three different languages (native Dutch, and non-native English and German).

Table 4.2. Estimated LE parameters from LGP experiments with group I subjects.

| Sentence type | Slope $\alpha$ | Offset $\beta$ | $R^2$ (explained variance) |
|---|---|---|---|
| Dutch (native) | –0.58 | 1.41 | 0.97 |
| English | –0.52 | 1.60 | 0.99 |
| German | –0.38 | 1.50 | 0.92 |

All differences between the values of $\alpha$ and $\beta$ in Table 4.2 are statistically significant. The influence of semantic constraints on LE, as quantified by slope $\alpha$, is as could be expected for group I: apparently, the semantic constraints present in German sentences are not used as effectively as in English sentences.

The differences in $\beta$ are not as easily interpreted, especially since $\beta$ is higher for English than for German. If we assume that $\beta$ expresses the linguistic entropy of words due to all factors *other* than semantic constraints,

then this also includes the systematic differences between orthographic representations of the different languages. In this light, the fact that $\beta$ is higher for English than for German does not seem as surprising anymore, but little room is left for interpretation of this parameter.

Table 4.2 shows that group I subjects benefit more from semantic constraints in English than in German. However, although it appears likely that there is a relation with speech intelligibility, Table 4.2 does not provide information about this relation.

By investigating similar curves as given in Fig. 4.9 for groups of subjects differing in (non-native) speech intelligibility, the relation between the $\alpha$-parameter and the SRT may be established.

For the data presented in Fig. 4.10, the 20 subjects of group I+II were divided in 4 subgroups according to their mean SRT when listening to German by G1 talkers. For these subgroups of 5 subjects, word-LE as a function of word position was calculated (Fig. 4.10 and Table 4.3).

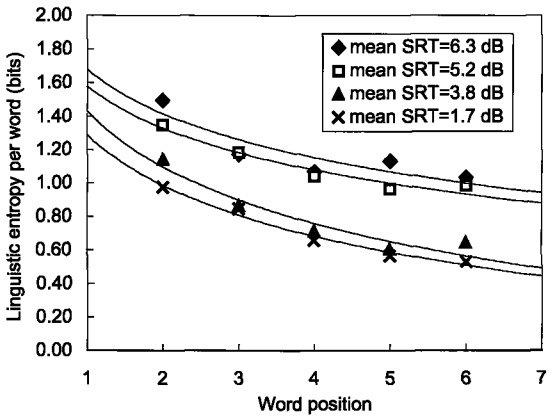


Figure 4.10. Non-native German word-LE as a function of word position within sentences, for word positions $2 \leq n \leq 6$. The dashed lines are least-squares fits of equation (3) to the data for 4 subgroups of subject group I+II, differing in mean SRT (G1 speakers). Data points are based on 5 subjects (each 39 sentences)

Table 4.3. Estimated LE parameters from LGP experiments with group I+II subjects (division into subgroups according to mean SRT scores for G1 talkers).

| Mean SRT of subgroup | Slope $\alpha$ | Offset $\beta$ | $R^2$ (explained variance) |
|---|---|---|---|
| 6.3 dB | –0.38 | 1.68 | 0.80 |
| 5.2 dB | –0.36 | 1.58 | 0.94 |
| 3.8 dB | –0.43 | 1.43 | 0.93 |
| 1.7 dB | –0.48 | 1.29 | 0.98 |

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

## 4.5. DISCUSSION AND CONCLUSIONS

Using the Speech Reception Threshold method, effects of non-native speech perception on speech intelligibility could be quantified for subjects ranging in proficiency from reasonable to excellent. Non-native speech recognition in noise does not just differ in terms of the mean of the psychometric function, but also the slope. To summarize the data given in this chapter, the average native (stylized) psychometric function and the worst-case non-native psychometric function derived from the experiments are given in Fig. 4.11.

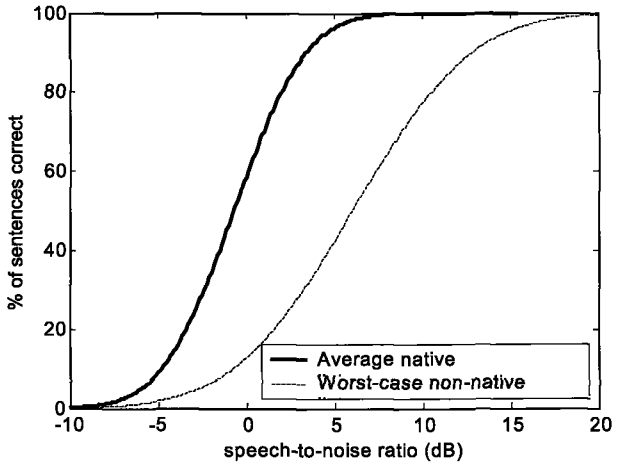


Figure 4.11. Psychometric functions of speech reception in noise (percentage of sentences correctly received as a function of speech-to-noise ratio) for the average native listener from the SRT experiments (SRT = –0.7 dB, steepness 12.6 %/dB) and the worst-case non-native listener (SRT = 6.0, steepness 7.5 %/dB).

The mean and slope of the psychometric functions of Fig. 4.11 can only be interpreted in the context of the specific sentence recognition paradigm used by the SRT test, implemented as described in Chapter 2. Other methods of measuring sentence recognition as a function of speech-to-noise ratio, or even other variations on the SRT paradigm, may lead to somewhat different results. For instance, relaxing the requirement that each individual word must

be responded correctly will reduce the steepness of the curve. On the other hand, if optimized sets of selected test sentences are used (Versfeld et al., 2000), then steeper psychometric functions will be found.

Despite the fact that there is a degree of dependence of the findings on the test method used, they also hold universal and quantitative meaning. If psychometric functions are known for two different test paradigms, in the same condition, then these curves can be used to transform measurement results from the scale of one test to the other. Hence, the difference between native and non-native intelligibility (given for our worst-case condition by the difference between both curves in Fig. 4.11) can also be transformed to other intelligibility scales, as long as the corresponding psychometric functions are known as a function of speech-to-noise ratio.

A non-native listener with a degree of command of his second language that is better than that of the worst-case listener presented in Fig. 4.11, will produce a psychometric function when subjected to an SRT test that is somewhere in-between the two curves of Fig. 4.11.

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

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

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

The experimental results offer no clear explanation for this discrepancy, but it seems that such an explanation is more likely to be found in the proficiency difference than in language-specific factors. The explanation could be that highly proficient listeners are able to use more subtle phonetic cues in authentically pronounced speech. The allophonic realizations of non-

native talkers, even if they match the listeners' native model of phoneme space better, are less effective in transferring information needed in the speech recognition process. For less proficient listeners, these subtle phonetic cues are not as useful; they are unable to accurately categorize allophones using typical L2 phonetic contrasts, and perform better if these L2 allophones are 'mapped' to their native phoneme space by non-native talkers.

In view of the results presented in Tables 4.2 and 4.3, it seems likely that the contradictory findings by Florentine (1985) and others versus Koster (1987), regarding the use of semantic constraints by non-native listeners, can be explained by differences in their test population's mean proficiency. A high proficiency population is likely to have 'near-native' use of contextual constraints, while this benefit is reduced for a low proficiency population.

It is important to note that none of the experiments presented in this chapter were concerned with subjects of very poor proficiency. The earliest stages of second language learning may involve intelligibility effects beyond our scope of interest. However, people with sufficient command of a second language for practical daily usage, will fall into categories somewhere between the two extremes given in Fig. 4.11. For the listener populations considered in this chapter, the presented measurement results can be used to assess exactly *where* between the lines in Fig. 4.11 we expect the psychometric function for a given population. For other languages and populations, additional data have to be collected. These data can consist of directly measured estimates of speech intelligibility; this is the best and most reliable option, but also the option that is the most difficult and time-consuming. Alternatively, listeners' intelligibility effects can be predicted from measures that are easier to obtain, such as linguistic entropy estimates.

# Chapter 5. Using the Speech Transmission Index for predicting non-native speech intelligibility[7]


## ABSTRACT

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. It is shown that the proposed correction function is also valid for conditions featuring bandwidth-limiting and reverberation. In the latter case, a non-standard range of modulation frequencies must be adopted in the STI calculations.


## 5.1. 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. In many cases where predictions of speech intelligibility are needed, the main interest is in the influence of the communication channel. The other factors are then (often implicitly) assumed constant. Objective speech intelligibility prediction 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

[7] This chapter is a slightly modified version of a manuscript submitted to J. Acoust. Soc. Am.: van Wijngaarden, S.J., Bronkhorst, A.W., Houtgast, T. and Steeneken, H.J.M. (subm). "Using the Speech Transmission Index for predicting non-native speech intelligibility."

Index (STI; IEC, 1998; STI; 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 Section 5.2, the suitability of existing objective speech intelligibility prediction models for non-native applications is discussed.

Section 5.3 continues by proposing a way in which the Speech Transmission Index (STI) can be used in various non-native scenarios. Section 5.4 contains a validation of this approach for speech in noise, bandwidth limiting and reverberation, in case of non-native listeners.

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

### 5.2.1. 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 at 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.

To predict the intelligibility of non-native speech, the interaction between speech transmission quality and language proficiency of talkers and listeners needs to be studied.

### 5.2.2. Features of the SII, STI and SRS models

At least three speech intelligibility prediction models presented in the 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. Comprehensive descriptions of these models are beyond the scope of this chapter. However, some features of each separate model that are related to suitability for non-native applications are summarized in this section.

#### *5.2.2.1. 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 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).

The STI is purely a measure of speech transmission quality: it indicates to what 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; however, a distinction is made between male and female speech. 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 5.1). The ranges of Table 5.1 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)

Table 5.1. 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 | - |

The STI is widely applied in room acoustics and telecommunications engineering. 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).

#### *5.2.2.2. The Speech Intelligibility Index (SII)*

The SII (ANSI, 1997) is an extension of a widely used version of the articulation index (Kryter, 1962), by incorporating 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 SII offers a choice between four calculation schemes, differing in frequency resolution (6 – 21 frequency bands). In cases where signals vary wildly with frequency, this allows the user to decide on a trade-off between complexity and accuracy.

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 an SII below 0.45, good communication yields an SII in excess of 0.75.

#### *5.2.2.3. 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; Müsch and Buus, 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 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.

#### *5.2.2.4. Comparison of STI, SII and SRS*

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, a non-native implementation of the SRS model is not pursued in this study. The main reason for this 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.

To incorporate effects of non-native speech into the STI or SII calculations, a possibility is to include a proficiency factor as suggested by Pavlovic et. al (1984). Errors in speech production and speech perception due to limitations of language proficiency can be used to calculate an appropriate value for this factor. This means that such an STI or SII value should always be accompanied by a detailed description of the talker and listener populations. Something similar always applies (even without using the proficiency factor) to the SII, since the SII depends on the type of intelligibility test it aims to predict.

For developing a practically feasible objective procedure for predicting non-native speech intelligibility, one needs to find an approach for incorporating knowledge on non-native speech communication into an existing prediction method. Such an approach, based on the STI, is outlined in the following section.

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

### 5.3.1. Rationale for correcting the qualification scale

Modifying the STI method by including a proficiency factor 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.

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 5.1) 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.

### 5.3.2. Method for correcting the qualification scale

#### *5.3.2.1. 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 (L1) and non-native (L2) 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 5.1 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 other speech degrading influences 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. 5.1 are fictitious. Intelligibility qualifications (Table 5.1) represent different levels of intelligibility (the vertical axis in Fig. 5.1). 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.

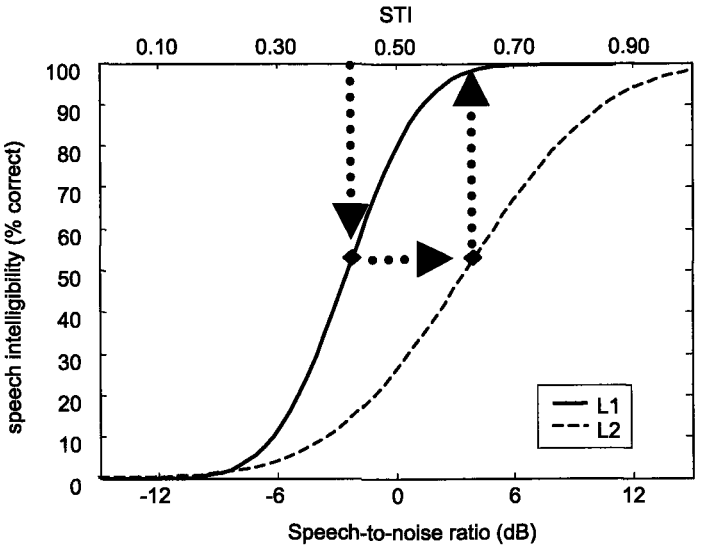


Figure 5.1. Schematic representation of the procedure for deriving a correction function for non-native interpretation of the STI. The native (L1) and non-native (L2) psychometric curves are fictitious, but representative of those found when measuring native and non-native sentence intelligibility.

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 operation visualized by Fig. 5.1 can only be carried out mathematically, if the relation $f(r)$ is reversible, meaning that Eq. 5.1 must be a unique function.

$$r = f^{-1}(\mathrm{STI}) \tag{5.1}$$

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. 5.1). The SNR then fully determines the STI, so each value of the STI corresponds to a single SNR[8]. All that is needed to calculate a correction function, is a model of the psychometric functions shown in Fig. 5.1. Of the possible choices for the independent variable $r$, the SNR is the easiest and most directly accessible option, and will be used throughout this chapter.

[8] 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.

After mathematically deriving (or numerically implementing) the correction of Fig. 5.1, it can be applied to the STI boundaries of Table 5.1. For each population of L2 talkers and listeners, the correction function will be different, leading to specific versions of Table 5.1.

#### *5.3.2.2. 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 be described by

$$\pi_{L1}(r) = \Phi\left(\frac{r-\mu_{L1}}{\sigma_{L1}}\right) \tag{5.2}$$

$\Phi(z)$ is the standardized cumulative normal distribution, $\mu_{L1}$ and $\sigma_{L1}$ are the mean and standard deviation of the distribution for fully native speech. A straightforward way to derive a correction function is to assume that Eq. 5.2 also holds for non-native speech, in which case $\mu_{L2}$ and $\sigma_{L2}$ will depend on the average proficiency level of the population. By solving $\pi_{L1} = \pi_{L2}$, substituting Eq. 5.1, a correction function as given in Eq. 5.3 is obtained.

$$STI_{L2} = f(\sigma_{L2}\left(\frac{f^{-1}(\mathrm{STI}_{L1})-\mu_{L1}}{\sigma_{L1}}\right)+\mu_{L2}) \tag{5.3}$$

Thus, assuming that, for a certain type of test that measures intelligibility as a function of $r$, $\mu_{L1}$ and $\sigma_{L1}$ are known, the information needed to correct a required $\mathrm{STI}_{L1}$ into an equivalent required $\mathrm{STI}_{L2}$, is a specification of the L2 population in terms of $\mu_{L2}$ and $\sigma_{L2}$.

Earlier results show that $\mu_{L2}$ and $\sigma_{L2}$, 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 (Fig. 4.11) as well as talkers (Fig. 3.6). In other words, L1 and L2 psychometric curves share a common origin (In Fig. 5.1 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 $\mu_{L2}$ and $\sigma_{L2}$, the L2 psychometric function can be derived from the L1 psychometric function using a single parameter $\nu$, according to Eq. 5.4.

$$\pi_{L2}(r) = 1-(1-\pi_{L1}(r))^{\nu} \tag{5.4}$$

The parameter $\nu$ (cf. the *j*-factor by Boothroyd and Nittrouer, 1988) can assume any value between 0 (no speech recognition at all) and 1 (native speech communication). It 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. 5.4, derived from an L1 psychometric function that follows a normal distribution, is shown in Fig. 5.2.

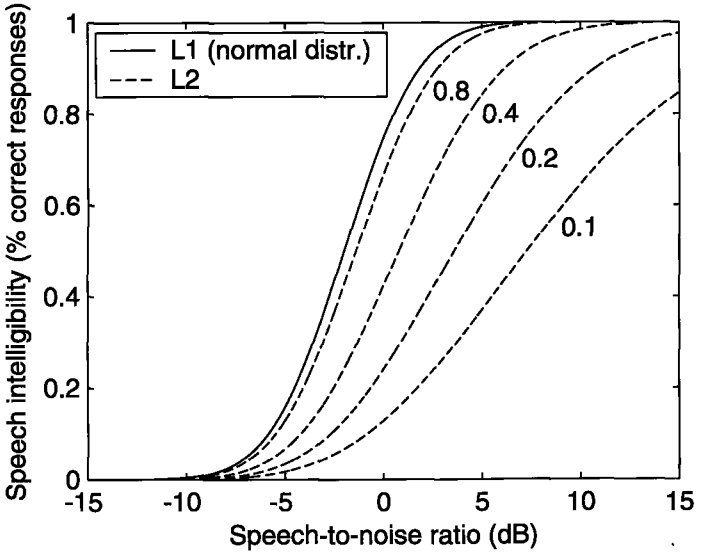


Figure 5.2. Examples of L2 psychometric functions derived from a cumulative normal L1 psychometric function ($\mu$ = –2, $\sigma$ = 3), according to Eq. 5.4, for $\nu$ = 0.8, $\nu$ = 0.4, $\nu$ = 0.2 and $\nu$ = 0.1.

It appears that Eq. 5.4 describes earlier experimental data very well, with only one parameter ($\nu$) instead of two ($\mu_{L2}$ and $\sigma_{L2}$), 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 $\mu$ and $\sigma$ 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. 5.3 can be noticeable.

A disadvantage of Eq. 5.4 is that a correction equation can not 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. 5.1). 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 Appendix A). 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 $v$ was easily realized, based on Eqs. 5.1, 5.2 and 5.4, following the procedure visualized in Fig. 5.1. This numerical implementation was used for calculating the correction functions used in this study.

#### *5.3.2.3. 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. 5.1 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 Wijngaarden et al., 2001a). The speech recordings that were used were part of the VU corpus (male talker) of SRT sentences (Versfeld et al., 2000).

### 5.3.3. Qualification labels for non-native listeners

#### *5.3.3.1. Correction functions for different populations of listeners*

Summarizing 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 (US 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. 5.3a. From the reported psychometric functions (given as Z-scores as a function of SNR), separate values of $\mu_{L1}$ and $\sigma_{L1}$ were taken for HP and LP sentences, and values of $v$ were obtained using a Gauss-Newton nonlinear fitting procedure. The correction functions for *HP* ($v = 0.36$) and *LP* ($v = 0.50$) sentences are given in Fig. 5.3b.

The difference between correction functions for high-predictability and low-predictability sentences is clear. The difference in $v$ can be seen as a quantification of Florentine's finding that non-natives are *not* as able as natives in making 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[9]. All groups consisted of 9 subjects. The original data are given in Fig. 5.4a, the derived correction functions in Fig. 5.4b.

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

[9] 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 is not used in this article.

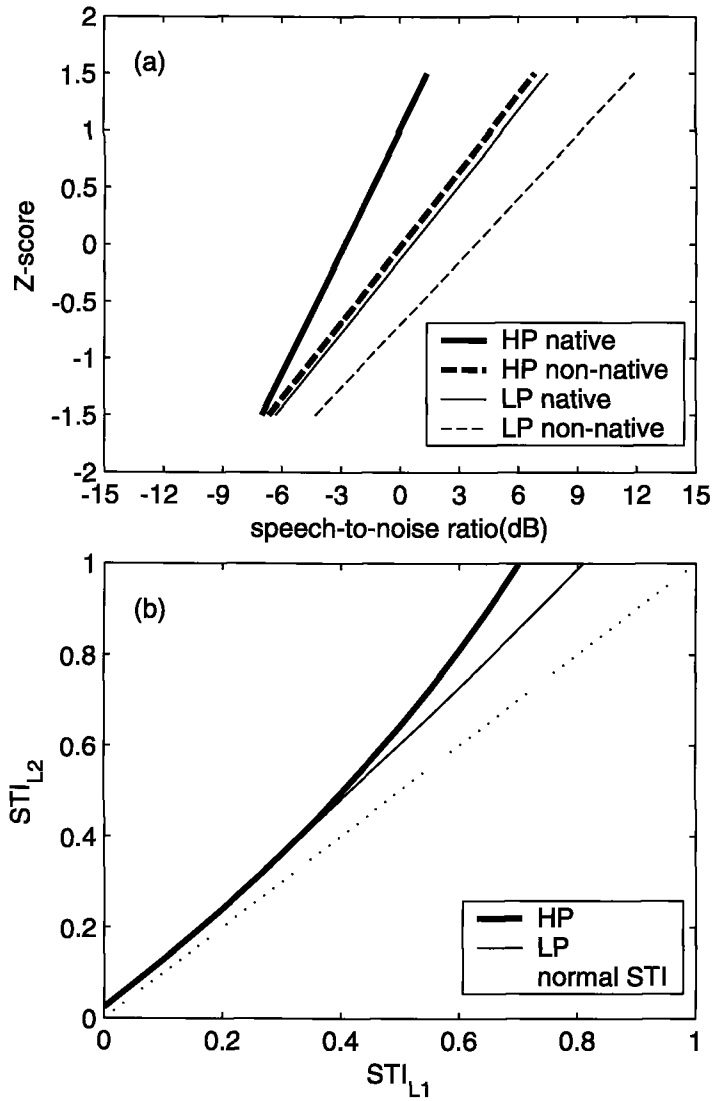


Figure 5.3. (a) Psychometric functions, in terms of $Z$-score as a function of SNR, for high-predictability (HP, $\nu$ = 0.36) and low-predictability (LP, $\nu$ = 0.50) sentences (after Florentine, 1985); (b) the STI correction functions derived from these psychometric functions.

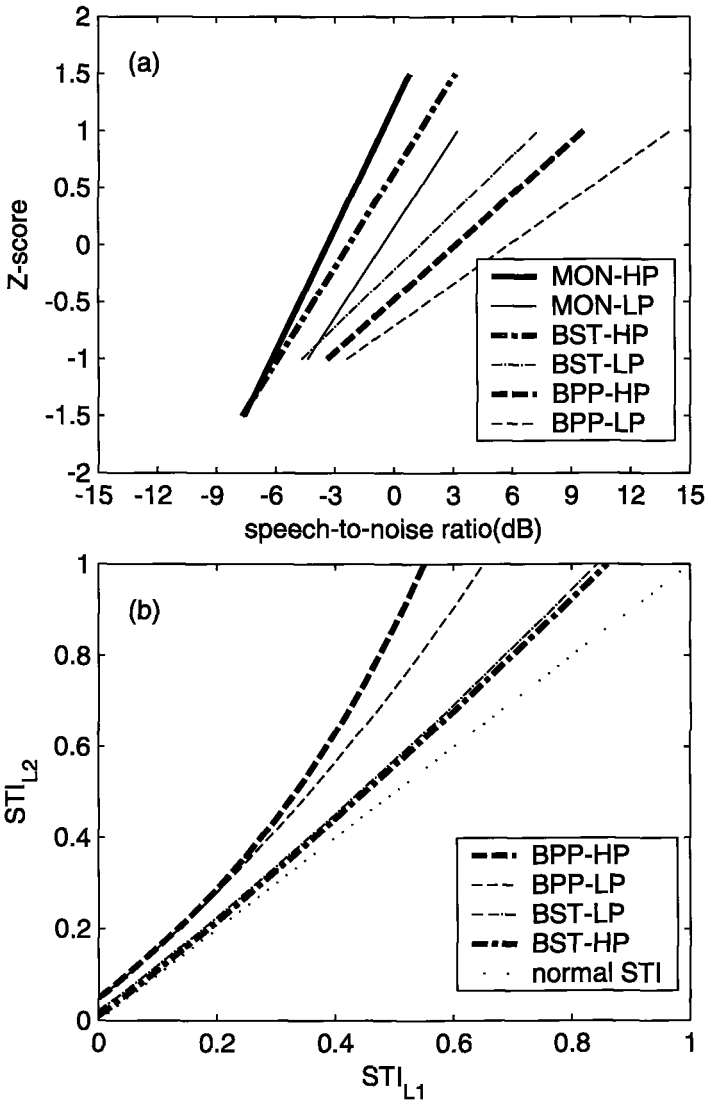


Figure 5.4. (a) Psychometric functions, in terms of Z-score as a function of SNR, for high-predictability (HP) and low-predictability (LP) sentences, for three groups of nine subjects: monolinguals (MON), early bilinguals (bilingual since toddler, BST) and late bilinguals (bilingual post puberty, BPP; after Mayo et al., 1997). (b) the STI correction functions derived from these psychometric functions.

Earlier data from tri-lingual non-native listeners (van Wijngaarden et al., 2002b) yield similar results for $\nu$-values and correction functions as the data by Mayo et al. The tri-lingual subjects were highly proficient in English, and

showed poor to moderate proficiency in German. The SRT sentence material used for obtaining these results are closest to the HP sentences of the SPIN test. Calculated mean $\nu$-values are 0.21 (German speech) and 0.52 (English speech). The corresponding STI correction functions are given in Fig. 5.5.

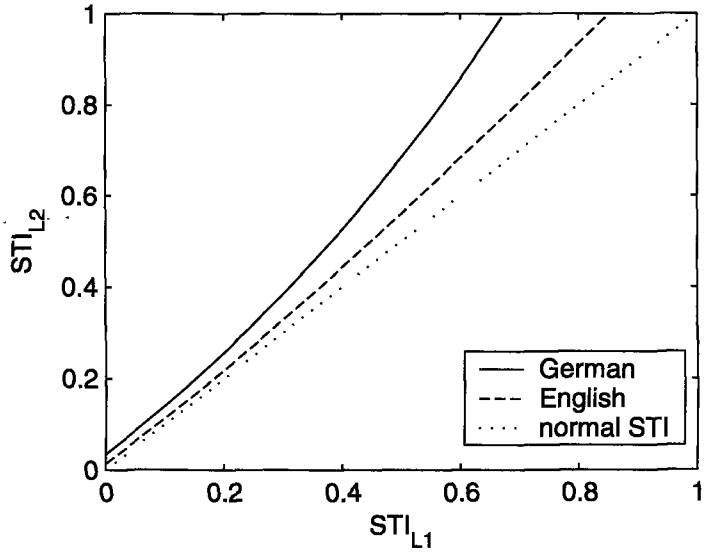


Figure 5.5. STI correction functions for trilingual Dutch listeners of German (low proficiency) and English (high proficiency). (after van Wijngaarden et al., 2002b)

### *5.3.3.2. Relation between STI and qualification labels for non-native listeners*

By applying the correction functions of Figs. 5.3, 5.4 and 5.5 to Table 5.1, the STI qualification label boundaries of Table 5.2 are obtained. From Figs. 5.3 and 5.4, the functions for HP sentences are used.

Table 5.2. Relation between STI and qualification labels for non-native listeners, after correction according to Figs. 5.3 and 5.4 (HP sentences), and Fig. 5.5. 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) | | Van Wijngaarden et al. (2002) | |
|---|---|---|---|---|---|---|
| | | | BST (early) | BPP (late) | English | 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 |

Table 5.2 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. 5.5 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 $\nu$-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 can not be related to the $\nu$-parameter on an individual level. However, the mean linguistic entropy $L$ can be compared to the mean value of $\nu$ 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 8 native and 8 non-native listeners. The non-native group consisted of L2 learners of the Dutch language, with different language backgrounds (Amharic, American English, German, Greek, Hungarian, Indonesian, Polish and Tigrigna) 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 and 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 having a 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 $\nu$-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 were 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. 5.4, by fitting the $v$-parameter.

A significant correlation was found between linguistic entropy and the $v$-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. 5.6, along with the means from the earlier experiments in German (low proficiency) and English (high proficiency).

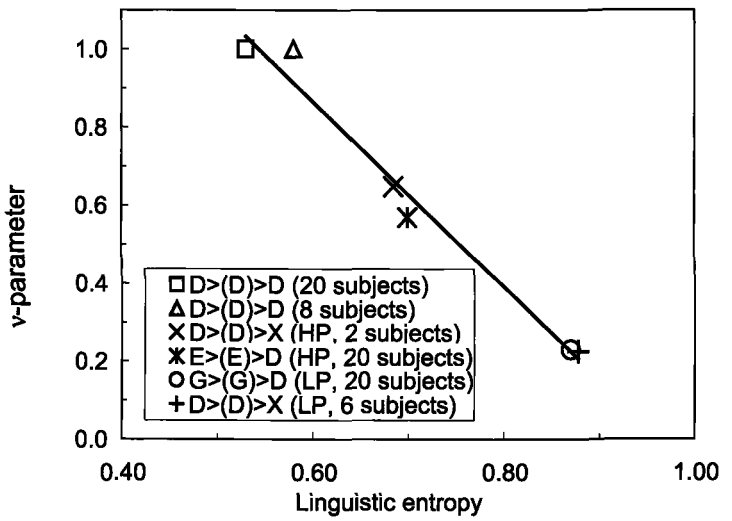


Figure 5.6. Relation between mean linguistic entropy and the $v$-parameter, for six groups of listeners: native Dutch listeners (D>(D)>D; a group of 20 and a group of 8 subjects), Dutch learners of English (E>(E)>D) and German (G>(G)>D), and two groups of learners of Dutch from various language backgrounds (D>(D)>X; 2 high proficiency listeners, 6 low proficiency listeners). The explained variance $R^2$ = 0.98.

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 $v$-parameter. The importance of this relation lies in the fact that the experimental procedures to determine a subject's linguistic entropy require only a fraction of the time needed to assess the $v$-parameter on an individual basis. Through the $v$-parameter, the interpretation of the STI for non-natives can be derived from linguistic entropy estimates.

### 5.3.4. 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[10]. The effect of a foreign accent on intelligibility can be predicted from ratings of accent strength, or from a talker's own opinion on the severity of his accent. Based on such measures, talkers can be categorized into accent strength categories (van Wijngaarden et al., 2002a; see Chapter 3). The four different categories of accent strength defined in Table 3.2 (numbered I-IV, ranging from 'native' to 'severe accent') were used to calculate STI correction functions (Fig. 5.7). The resulting STI label categories are given in Table 5.3.

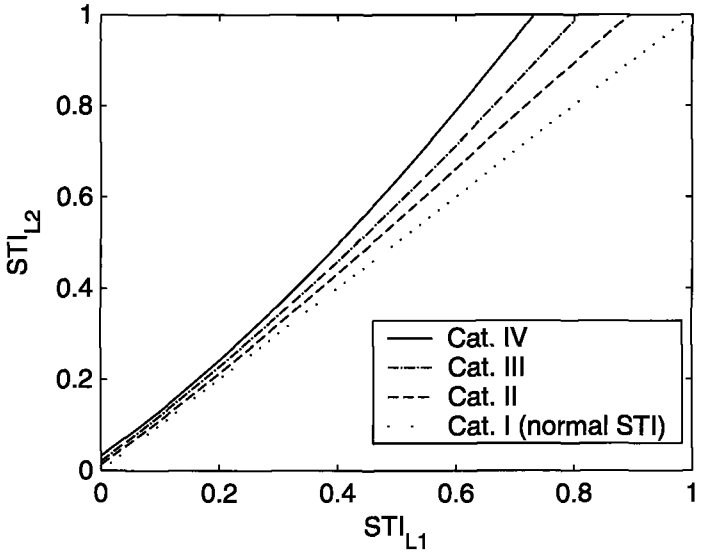


Figure 5.7. STI correction functions for L2 talkers of the Dutch language, for different degrees of foreign accent strength (cat. I–cat. IV; van Wijngaarden et al., 2002a). Category I means that the talker has (virtually) no foreign accent, category IV means that the accent is severe (see Table 5.3 for the corresponding values of $v$).

[10] This statement is based on comparisons of SRT results between cases where only the talker is non-native, or 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.

Table 5.3. Relation between STI and qualification labels for non-native talkers differing in degree of foreign accent, after correction according to Fig. 5.7. The text ">1" indicates that an STI greater than 1 would be required, meaning that this qualification cannot be reached. The mean $v$-value for each category is also given.

| STI label category boundary | Standard STI (Cat. I) | Cat. II ($v$ = 0.67) | Cat. III ($v$ = 0.48) | Cat. IV ($v$ = 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 |

Figure 5.7 and Table 5.3 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 5.2 and 5.3.

## 5.4. VALIDATION OF THE QUALIFICATION SCALE CORRECTION

### 5.4.1. Validation issues

If speech is degraded by additive noise only, there seems 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 noise, the validity of this approach remains to be proven. Two important sources of speech degradation are bandwidth-limiting and reverberation.

All measurements described so far were based on wide-band conditions with an equal SNR at each frequency. The STI method takes frequency range effects on intelligibility into account by analyzing the SNR in seven separate octave bands (125 Hz–8 kHz), using a weighting function for the relative differences in importance between octave bands. In a relatively recently standardized version of the STI (IEC, 1998; Steeneken and Houtgast, 1999), which is used throughout this paper, the frequency-dependent relation between the STI and native speech intelligibility is improved by taking neighboring octave band dependence into account. 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 this MTF-based operation is just as 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. 5.1). However, the only independent parameter ($r$ in Eq. 5.1) 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 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.

### 5.4.2. Effects of bandwidth limiting

The same 16 listeners who participated in the SRT and LGP experiments reported above and shown in Fig. 5.6, 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 band-limited speech, to prevent spreading of information into adjacent bands through non-linear 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 of the fact that 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. Results are given in figure 5.8.

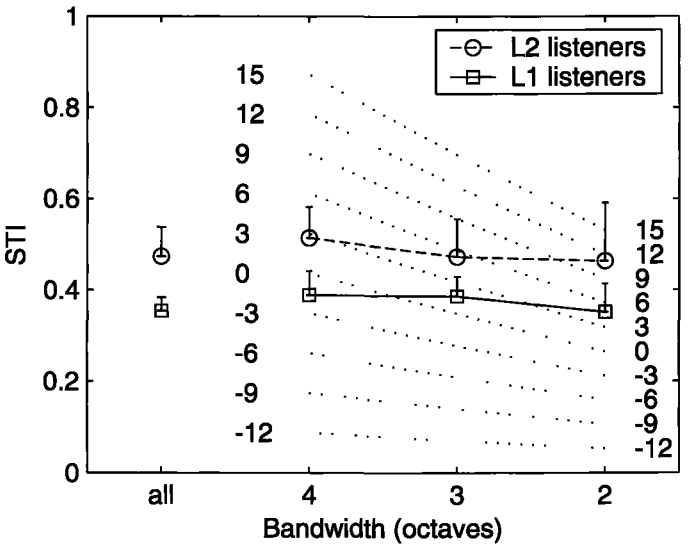


Figure 5.8. STI at the SRT, for conditions with and without bandwidth-limiting. The dotted lines indicate the maximum STI at each bandwidth, as a function of the SNR. The error bars indicate the standard deviation ($N = 24$; 8 listeners, each 3 SRT measurements per condition).

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 3-octave condition for the native group, none of the within-group differences in Fig. 5.8 are 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.

The mean native STI-results fall in the range between 0.30 and 0.45, leading to a classification "poor" according to the standard table (Table 5.1). The mean non-native results for each condition would be (incorrectly) categorized as "fair."

The $\nu$-value for each non-native listener was determined in a separate experiment, following the procedure described above in relation to Fig. 5.6. Using the mean value of the $\nu$-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 $0.37 < \text{STI} < 0.59$).

### 5.4.3. 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.

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 approximately 1.5 seconds 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 pre-recorded 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 approximately 0.5 and 2 seconds 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' (2 subjects) and 'low-proficiency' (6 subjects) used earlier, was again applied. Results of STI calculations at the SRT as a function of EDT, similar to Fig. 5.8, are given in Fig. 5.9.

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 both normal hearing and hearing impaired listeners (Duquesnoy and Plomp, 1980). If the standard procedure for calculating the STI (based on a modulation frequency range of 0.63–12.5 Hz) had been applied for obtaining Fig. 5.9, the effect of reverberation would have been underestimated. This problem, related to speaking style and envelope spectrum of the talker, will be addressed in Chapter 6. The STI calculations underlying Fig. 5.9 are based on a modulation frequency range of 0.63–31.5 Hz.

For all three groups in Fig. 5.9, 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.

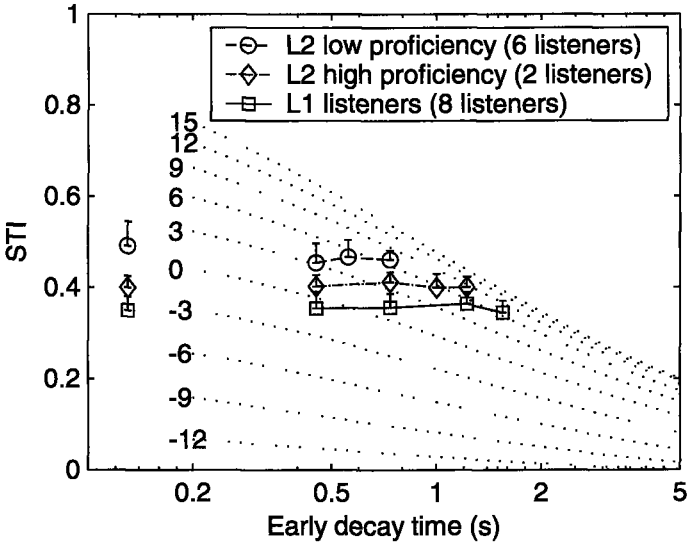


Figure 5.9. STI at the SRT, for conditions with and without reverberation. The dotted lines indicate the maximum STI at each EDT, as a function of the SNR. The error bars indicate the standard deviation (2 to 8 listeners, each 3 SRT measurements per condition). The EDT in this plot is the mean EDT in the octave bands 125 Hz–8 kHz. The STI calculation is non-standard, and includes modulation frequencies up to 31.5 Hz.

## 5.5. DISCUSSION AND CONCLUSIONS

### 5.5.1. The $\nu$-parameter

The approach for non-native interpretation of the STI, as proposed in this thesis, is based on a few novel concepts. Perhaps the most important of these is modelling the non-native psychometric function by relating it to the native psychometric function, through a single parameter $\nu$. 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 psychometric curves (such as seen in Fig. 3.6) matches expectations based on differences in the $\nu$-parameter. This leads us

to conclude that this non-native psychometric function model is the most appropriate choice for our current purposes.

### 5.5.2. Effects of linguistic message content

Our correction function approach yields, by definition, representative results if the only speech-degrading factor is additive 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 5.4 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. 5.3 and 5.4). 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, then the specific details of linguistic content are considered to be of minor importance. In Figs. 5.3 and 5.4, the HP curves are expected to be most representative of the STI application domain.

### 5.5.3. STI modulation frequency range

In the calculations of the STI under the influence of reverberation, the range of modulation frequencies included in the STI calculations was extended to 31.5 Hz. This was done because of considerations related to the specific characteristics of the talker. This problem has not been properly dealt with here, but will be addressed in the next chapter.

### 5.5.4. 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 chapter 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 25$^{th}$ percentile, in which case 75% of the population shows equal or better proficiency than the threshold). Using the relation shown in Fig. 5.6, the chosen entropy threshold can be converted into the equivalent value of the $\nu$-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. 5.7, self-ratings can also be translated into equivalent values of the $\nu$-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.

# Chapter 6. Effect of talker and speaking style on the Speech Transmission Index[11]


## ABSTRACT

The Speech Transmission Index (STI) is routinely applied for predicting the intelligibility of messages (sentences) in noise and reverberation. Despite clear evidence that the STI is capable of doing so accurately, recent results indicate that the STI sometimes underestimates the effect of reverberation on sentence intelligibility. To investigate the influence of talker and speaking style, the Speech Reception Threshold in noise and reverberation was measured for three talkers, differing in clarity of articulation and speaking style. For very clear speech, the standard STI yields accurate results. For more conversational speech by an untrained talker, the effect of reverberation is underestimated. Measurements of the envelope spectrum reveal that conversational speech has relatively stronger contributions by higher (>12.5 Hz) modulation frequencies. By modifying the STI calculation procedure to include modulations in the range 12.5–31.5 Hz, better results are obtained for conversational speech. Envelope spectra were also measured for a population of 134 randomly selected talkers, revealing that the differences among the three talkers used for the present SRT experiments are representative of those encountered among the population.


## 6.1. INTRODUCTION

The Speech Transmission Index (Steeneken and Houtgast, 1980) is a physical measure for objectively predicting the intelligibility of speech. The Speech Transmission Index (STI) model, built on the general concepts of the Articulation Index (French and Steinberg, 1947; Kryter, 1962), uses modulation transfer functions (MTFs) to predict intelligibility under the influence of a wide diversity of speech degradations. Throughout the decades, the STI was developed from a collection of ideas related to the

[11] This chapter is a slightly modified version of a manuscript submitted to J. Acoust. Soc. Am.: van Wijngaarden, S.J. and Houtgast, T. (subm). "Effect of talker and speaking style on the Speech Transmission Index."

speech envelope spectrum (Houtgast and Steeneken, 2002) into a standardized and widely applied evaluation tool (IEC, 1998; ISO, 2002).

One of the attractive features of the STI model, especially for applications related to room acoustics, is its ability to predict how degradations of temporal properties of the speech signal, as caused by reverberation and echoes, reduce speech intelligibility (Houtgast et al., 1980). Through the modulation transfer function, these influences are translated into "equivalent speech-to-noise ratios," and then treated in essentially the same way as additive noise.

The STI method was designed and optimized to yield representative and homogeneous intelligibility predictions across all kinds of speech degradation. A certain reduction in (subjective) intelligibility corresponds to a matching reduction in STI, whether this intelligibility reduction is due to noise, reverberation, peak clipping or something else. This was validated by means of subjective experiments using consonant-vowel-consonant (CVC) words across wide ranges of test conditions, including noise and reverberation (Steeneken and Houtgast, 1980). Hence, at least for CVC words, the effects of noise as well as reverberation are incorporated with equal weight.

Using the speech reception threshold (SRT) method (Plomp and Mimpen, 1979), the same was found for short, redundant sentences (Duquesnoy and Plomp, 1980). The SRT is expressed as the speech-to-noise ratio corresponding to 50% sentence intelligibility. Duquesnoy and Plomp measured the SRT as a function of early decay time (EDT), and then evaluated the STI at this speech-to-noise ratio (the "STI at the SRT"). It was shown that 50% sentence intelligibility always corresponds to the same STI, whether noise is the predominant speech degrading factor, or reverberation.

Recent results (reported in Chapter 5, using an adapted STI calculation procedure) show that this is not always the case. The experiment of Duquesnoy and Plomp was essentially repeated, this time with native as well as non-native listeners, but with a different set of speech recordings (Versfeld et al., 2000). In this case, the standard STI method was found to have a tendency to underestimate the effect of reverberation on sentence intelligibility. A similar mismatch between subjective intelligibility and the STI in combined "noise plus reverberation" conditions has been reported before (Payton et al., 1994; Fig. 10, triangular data points on the left). The mismatch reported by Payton et al. (1994) seems to depend on speaking style, and is larger for a conversational than for a clear speaking style. This suggests that the difference may be due to the speech recordings that were used.

In the next section, experiments along the lines of Duquesnoy and Plomp (1980) are described, in which the effect of noise and (simulated) reverberation on the STI corresponding to 50% sentence intelligibility is studied. By evaluating the interaction with different speech materials, the influence of talker-specific characteristics (mainly clarity of articulation and

speaking style) is addressed. These experiments were carried out with native and non-native listeners. In Section 6.3, an explanation (and remedy) for the discrepancies in the results reported in Section 6.2 is offered.

The version of the STI method used throughout this chapter is the revised STI ($STI_r$), based on the most recent version of the standard available at this time (IEC, 1998).

## 6.2. SENTENCE INTELLIGIBILITY IN NOISE AND REVERBERATION

### 6.2.1. Method

The speech reception threshold (SRT; Plomp and Mimpen, 1979) is the speech-to-noise ratio at which 50% intelligibility of short, redundant sentences is realized. The SRT is measured using an adaptive up-down procedure. The original corpus of speech recordings made by Plomp and Mimpen (10 lists of 13 Dutch sentences, uttered by a female talker in a very clear voice) has seen extensive application, among which the experiments of Duquesnoy and Plomp (1980). They were also used in the present study.

A new, much larger, corpus of SRT test sentences is the "VU" corpus (Versfeld et al., 2000). This corpus (also in Dutch) consists of 39 lists of 13 sentences of a male talker, and the same amount of material for a female talker. The sentences by the male talker were used in this experiment. Versfeld et al. present the VU sentences as roughly equivalent to the Plomp and Mimpen sentences. However, the author of this thesis perceives the adopted speaking style to be less clear.

A third corpus of SRT sentences is the multi-lingual SRT (ML-SRT) database (van Wijngaarden et al., 2001a; van Wijngaarden et al., 2002b). This corpus consists of material by many talkers in various languages. These talkers are, by contrast to the Plomp and Mimpen and VU sentences, non-professionals; they have had no experience or training as actors, singers or announcers. The adopted speaking style differs slightly between talkers in the ML-SRT corpus. The single male Dutch talker used in the present study speaks less clearly than the VU talker, and certainly less clearly than the Plomp and Mimpen talker.

The masking noise used in the SRT procedure was noise with the same long-term spectrum as speech by the corresponding talker (as routinely provided with all speech material intended for SRT testing). Reverberation can be created or simulated in a number of ways. In earlier experiments, use was made of pre-recorded impulse responses corresponding to various Early Decay Times (EDT). Noise was mixed with the target speech samples, after which this signal was convolved with a suitable impulse response to recreate reverberant speech in noise. This approach, theoretically almost equivalent to monaural listening in a real reverberant environment, was not adopted in the

current experiments. When manipulating the EDT in a room by adding or removing acoustic absorption, the timbre of sounds in the room is also changed slightly, in a non-systematic fashion. To exclude effects such as these, synthetic impulse responses were created by subjecting white noise to pure exponential decay. Such impulse responses of various EDTs, with a length of 1–5 s depending on EDT, were convolved with target speech to obtain pseudo-reverberant speech. In terms of the modulation transfer function, the resulting pseudo-reverberant conditions are identical to purely exponentially decaying real reverberant conditions of the same EDT. The applied pseudo-reverberation may appear somewhat artificial, at least to acoustically educated listeners, but offers maximum control over the MTF.

### 6.2.2. Native listeners

Figure 6.1 shows "STI at the SRT" results based on the three different speakers, obtained with eight native Dutch listeners. First, the individual SRT was measured in a number of reverberation conditions. From this, the STI at the SRT (the STI corresponding to 50% sentence intelligibility) was calculated. If the STI model predicts effects of reverberation as accurately and unbiased as effects of noise (when related to sentence intelligibility), then the lines in Fig. 6.1 are straight and horizontal.

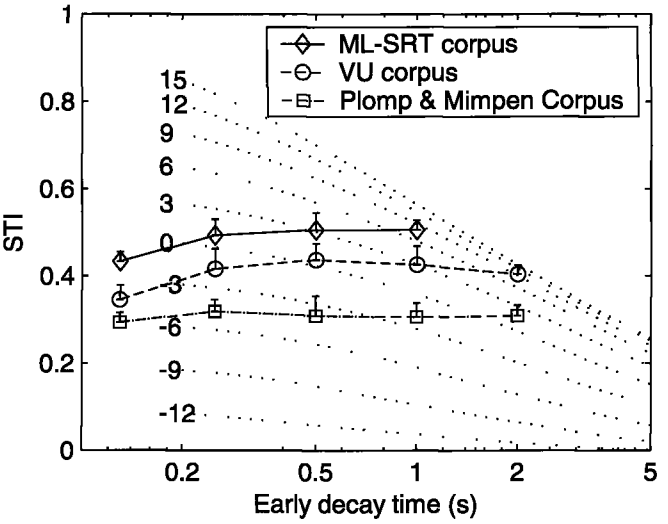


Figure 6.1. STI at the SRT (native Dutch listeners), for conditions with and without synthetic reverberation. The dotted lines indicate the maximum STI at each EDT, as a function of the SNR. The error bars indicate the standard deviation (8 listeners, each 2 SRT measurements per condition). The leftmost data point of each line represents a condition without reverberation.

The three talkers represented in Fig. 6.1 differ in terms of their average intelligibility; the three lines differ significantly. The Plomp and Mimpen talker offers the highest intelligibility, the ML-SRT talker the lowest.

Figure 6.1 clearly shows that 50% sentence intelligibility sometimes corresponds with a higher STI *with* than *without* reverberation. For the Plomp and Mimpen talker, the line in Fig. 6.1 follows the theoretical straight and horizontal line. There is a significant difference only between the STI without reverberation and the STI at EDT = 0.25 s, but this difference is relatively small. This essentially replicates the results found by Duquesnoy and Plomp (1980). For the VU and ML-SRT talkers, there is a mismatch; the STI without reverberation differs significantly ($p < 0.05$) from the STI in any reverberation condition. This is in agreement with results reported in Chapter 5, this time for synthetic reverberation.

Figure 6.1 shows that using the STI for predicting the intelligibility of sentences under reverberant conditions may or may not result in errors, depending on the combination of talker and speaking style. It does not explain *why* talker and/or speaking style make a difference. This is addressed in Section 6.3.

### 6.2.3. Non-native listeners

To allow interpretation of the STI for non-natives with a simple correction function, the STI at the SRT needs to be independent of reverberation times, or else an error is made. This must be true in case of native as well as non-native SRT measurements. To investigate how Fig. 6.1 translates to a non-native scenario, a similar experiment was carried out with 8 non-native listeners of varying Dutch proficiency (see Chapter 5 for a description of this subject population; mean $v$=0.33). Results are reported in Fig. 6.2.

Compared to Fig. 6.1, all lines in Fig 6.2. are shifted to higher STI values. Again, the desired horizontal line is closely approximated for the Plomp and Mimpen talker. Significant differences are found for the VU and ML-SRT talker, but not for the Plomp and Mimpen talker.

The fact that the clear speech of the Plomp and Mimpen talker leads to an STI at the SRT that is independent of EDT, for native as well as non-native listeners, means that for this type of speech the correction functions proposed in Chapter 5 can be applied with the standard STI method without further consideration.

At first sight, the effect of EDT on the STI at the SRT, as shown in Figs. 6.1 and 6.2, may appear relatively small. However, even a small (but systematic) mismatch between the way that the STI is affected by noise versus reverberation may cause considerable complications when using the STI in practice.

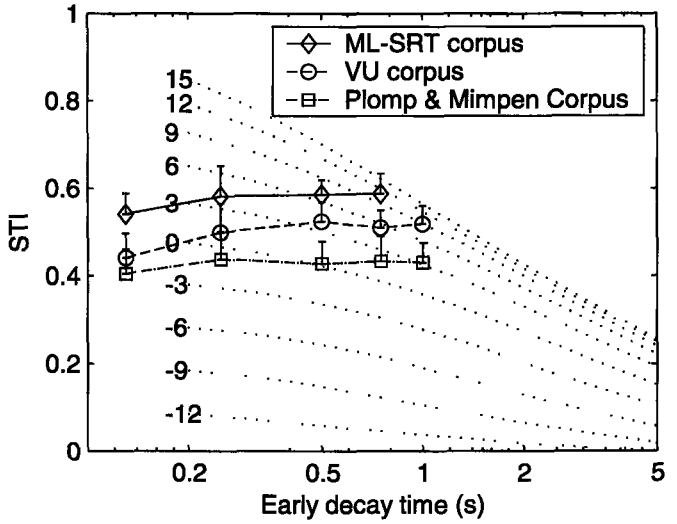


Figure 6.2. STI at the SRT (non-native listeners), for conditions with and without synthetic reverberation. The mean value of the ν-parameter for this population of non-native listeners is 0.33

## 6.3. EXPLANATION FOR THE EFFECT OF SPEAKING STYLE

### 6.3.1. Trends observed in the data

Figs. 6.1 and 6.2 show that, under reverberant conditions, the STI is overestimated even at low reverberation times. Even a small amount of reverberation has an impact on intelligibility, to a degree not predicted by the STI model. This effect only appears for talkers adopting a more informal, conversational speaking style; the effect is absent for the Plomp and Mimpen talker.

These observations can be explained by assuming that the relation between the envelope spectrum of speech and intelligibility depends on speaking style. The way that the STI model relates intelligibility to the modulation transfer function is apparently quite suitable for some talkers (and speech styles), but less so for others. This thought is explored further in the remainder of this section.

The observations could also be explained in a number of other ways, such as the assumption that the relevant range of speech-to-noise ratios is not always, as assumed by the STI model, covered by a linear 30-dB range, where each 1-dB increase has an equal effect on the STI. Depending on the individual talker or speaking style, a specific "intensity importance function" (Studebaker and Sherbecoe, 2002) could be adopted to predict the observed differences. However, the data in Fig. 6.1 do not appear very consistent with such an explanation.

### 6.3.2. Between-corpus differences in the speech envelope spectrum

The STI model uses a fixed (logarithmic) set of 14 modulation frequencies ranging from 0.63 to 12.5 Hz, at $^1/_3$-octave intervals. All modulation frequencies have equal weight. This represents, more or less, the modulation frequency range observed in natural speech. The envelope spectrum of speech normally shows a maximum around 3 Hz, and contains almost all of its energy in the range from 0–30 Hz.

The modulation frequency range in the STI model, and the choice of uniform weighting of the modulation frequencies, are design choices based on the desired correspondence between the effect of noise and reverberation on the STI. For the chosen range, good correspondence between intelligibility of CVC nonsense words and the STI was observed (Steeneken and Houtgast, 1980). As shown above, this good correspondence sometimes holds for short sentences, but apparently only for clear speech by a trained talker. If differences in clarity of articulation and speaking style translate into differences in the envelope spectrum, something is to be said for adopting different modulation frequency weighting schemes for different speaking styles.

Envelope spectra were calculated from the recorded SRT sentences. Sentences were concatenated, separated by silences of a random duration up to 500 ms, to form sequences of approximately 30 seconds in length. Modulation spectra were calculated for ten of these sequences per talker, taking the average across these sequences to obtain a more accurate estimate of the envelope spectrum.

The method for calculating envelope spectra essentially follows the procedure originally proposed in the context of the STI model (Houtgast et al., 1980), but is implemented in digital algorithms rather than analog hardware. The speech (sampled at 44,100 Hz) is band-filtered into the seven audio-frequency octave bands used by the STI model. Next, the signal is squared, down-sampled by a factor of 300, repeated ten times to obtain a better frequency resolution, and then subjected to a discrete Fourier transform. The obtained line spectrum is normalized by its DC-component to allow interpretation in terms of modulation-indices, and binned into 1/3 octave bands in the range from 0.40 to 31.5 Hz. This gives a separate modulation spectrum for each of the audio-frequency octave bands.

As shown in Fig. 6.3, the envelope spectra for the three different speech materials all have the usual maximum around 3–4 Hz, but show differences in magnitude. Results appear different for each individual audio frequency octave band, making it difficult to detect systematic differences due to the speech material.

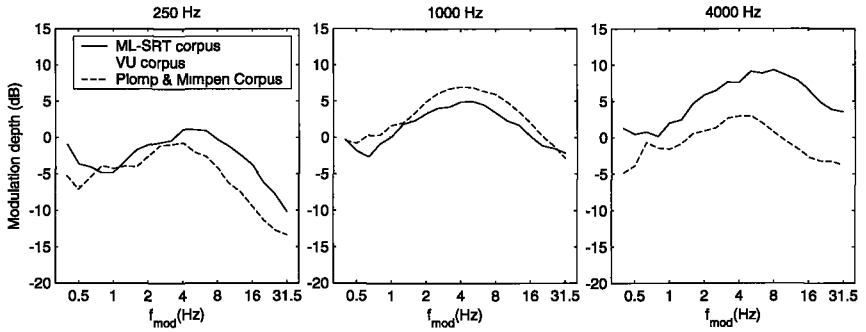


Figure 6.3. Averaged envelope spectra (250, 1000 and 4000 Hz audio frequency bands) of speech by three different talkers.

By inspecting envelope spectra such as Fig. 6.3, the only (subtle) trend that may be observed, is that for the clearer Plomp and Mimpen sentences, the energy in the envelope spectrum appears to be concentrated more around the maximum at 3 Hz. It spreads a smaller fraction of its total energy to higher modulation frequencies.

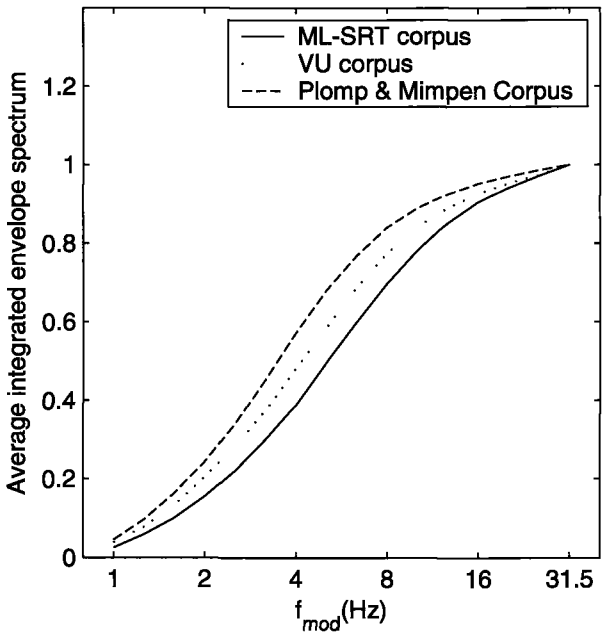


Figure 6.4. Integrated (cumulative) envelope spectra of speech by three different talkers. The square of modulation index $m$ was integrated from 1 Hz upward, and averaged across the audio frequency octave bands 125–8000 Hz. The integrated spectrum was normalized to unity at 31.5 Hz.

To investigate whether this is a systematic effect, frequency-integrated versions of the envelope spectra were calculated, averaged across audio frequency and normalized by dividing through their cumulative maximum (making the value at 31.5 Hz, the highest measured modulation frequency, equal to 1). Figure 6.4 shows these integrated (or cumulative) spectra for the three different speech materials, integrated from 1 Hz upward.

The tendency in Fig. 6.4 appears to be that the envelope spectrum of clearer speech show relatively smaller contributions by the higher modulation frequencies. The modulation frequencies in Fig. 6.4 not taken into account by the STI model (> 12.5 Hz) represent only a small portion of the total energy for the Plomp and Mimpen corpus, but are of greater importance for the ML-SRT and VU material.

### 6.3.3. Adapting the STI method by using a wider modulation frequency range

A straightforward first step in trying to adapt the STI model to more conversational speech would be to extend the modulation frequency range to 31.5 Hz, maintaining equal weight for all modulation frequencies. The modulation frequencies remain separated by 1/3 octave, so the extension to 31.5 Hz increases the number of modulation frequencies from 14 to 18 [12].

Figure 6.5 is based on the same SRT data as Fig. 6.1, this time with the modulation frequency range for the STI calculation extended to 31.5 Hz. The ML-SRT data in Fig. 6.5 shows a much closer resemblance to the expected horizontal line than in Fig. 6.1. The same is true for the VU data, even if some dependence of the STI on the EDT is still observed (the STI at EDT=0.50 differs significantly from the STI without reverberation). Only for the Plomp and Mimpen data, Fig. 6.1 fits the expected horizontal line better. This confirms the expectations based on the modulation spectra of Fig. 6.4.

[12] Extension of the range to higher modulation frequencies is one of the ways in which the STI model can be made more sensitive to reverberation. Another possibility would have been to maintain 14 modulation frequencies, but shift the entire range upward. It has been verified that, for the present data, this leads to essentially similar results. However, this would also affect the relation between the STI and intelligibility for conditions affecting the low-frequency end of the envelope spectrum, such as AGC (automatic gain control).

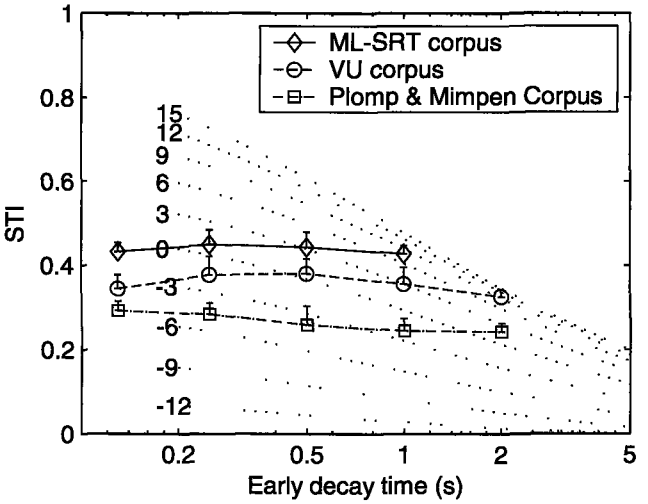


Figure 6.5. STI at the SRT results, based on the same data as Fig. 6.1, but with a non-standard modulation frequency range (0.63–31.5 Hz). The dotted reference lines (STI vs. EDT as a function of SNR) are also based on this extended modulation frequency range.

### 6.3.4. Envelope spectra for a larger population of talkers

Given the differences in modulation spectra for the three SRT talkers, the question rises what variations may be expected for a greater population of (arbitrarily selected) talkers. Subjective impressions indicate that the ML-SRT talker and the Plomp and Mimpen talker represent opposite ends of the clarity range for normal, non-pathological speech free of "mumbling". Do their modulation spectra also represent opposite ends of the range?

Speech material (Dutch newspaper sentences) read aloud by 134 (male and female) native Dutch talkers, taken from the Dutch NRC corpus (having a similar structure as the Wall Street Journal corpus, Paul and Baker, 1992; van Leeuwen and Orr, 2000), was subjected to the same modulation spectrum calculations as the SRT sentences. The NRC corpus consists of high quality recordings of untrained talkers, screened for impairments, but otherwise randomly selected.

Figure 6.6 shows that the maximum of the envelope spectrum shifts slightly (from approx. 3 to 4 Hz) for higher audio frequencies. The statistical spread is considerable, especially for the higher frequency bands. The bottom right panel of Fig. 6.6 shows 5$^{th}$–95$^{th}$ percentile versions of the integrated envelope spectrum, derived from the individual talker data rather than by integrating the curves shown in the other panels. The 5$^{th}$ and 95$^{th}$ percentile curves in this panel are close to the ML-SRT and Plomp and Mimpen data, respectively, in Fig. 6.4. This indicates that the Plomp and Mimpen and

ML-SRT talkers represent the extremes of the talker population on which Fig. 6.6 is based.

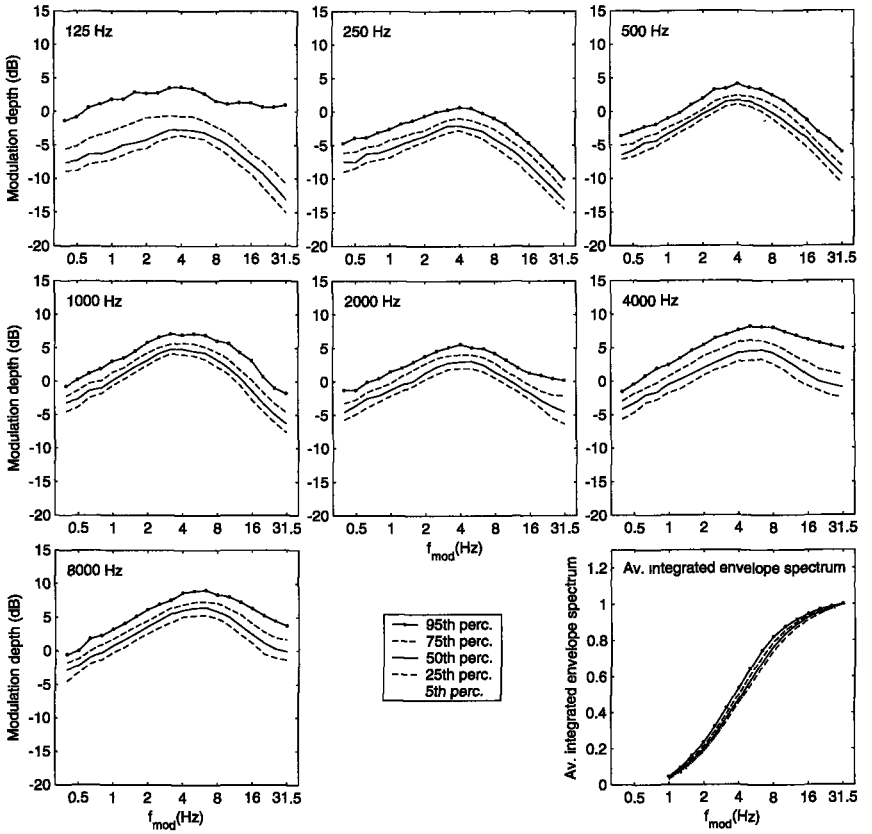


Figure 6.6. Envelope spectra (125–8000 Hz audio frequency octave bands) of speech by 134 different talkers. The data is represented by the 5th, 25th, 50th, 75th and 95th percentile. The corresponding integrated (cumulative) envelope spectra (the square of modulation index *m* integrated from 1 Hz upward and averaged across all audio frequency octave bands) are also given.

The overall energy in the envelope spectra in Fig. 6.3 is audio-frequency octave band-dependent: the order of the curves in the 1000 Hz panel is exactly the opposite of those in the 4000 Hz panel. Inspection of the data behind Fig. 6.6 reveals no systematic correlation supporting a more general trend along the lines of the reversal in Fig. 6.3. This suggests that the reversal between 1000 Hz and 4000 Hz in Fig. 6.3 is coincidental.

## 6.4. CONCLUSIONS AND DISCUSSION

Depending on the talker and the adopted speaking style, the standardized STI calculation procedure (IEC, 1998) may give inaccurate predictions of sentence intelligibility in reverberant conditions. Based on the data presented in this paper, we propose to apply a wider range of modulation frequencies (0.63–31.5 Hz instead of 0.63–12.5 Hz) for predicting the intelligibility of conversational speech. For clear speech, the standard modulation frequency range is more appropriate.

Further fine-tuning of the STI model may be possible through the application of modulation frequency weighting functions. For the limited range of variations in speaking style and voice quality addressed in this study, such a refined approach is not necessary. However, for more extreme variations in speaking style (including true conversations, where the interaction between the communicators becomes important) the STI model may benefit from a more refined approach involving modulation frequency weighting functions.

# Chapter 7. General discussion

The problem of cross-language speech communication is highly complex. It is a large and challenging puzzle, which has been keeping linguists, phoneticians, and scientists from other speech-related disciplines busy for many decades. This study, like any other, can provide only a limited contribution to solving the puzzle, and only from a very specific perspective.

As stated in the introduction to this thesis, the perspective of this study is application-oriented. This sets the study apart from others: most cross-language speech perception research is initiated to satisfy our scientific interest in the complicated processing underlying human speech communication. Non-native speech gives away some of the secrets of the complex perceptual processing underlying speech communication, which remain hidden in 'normal' speech. For instance: perceptual confusions among phonemes for second-language learners may help us understand the process of phonetic categorization. But how will this knowledge help scientists, system designers and engineers who are faced with "the non-native factor" as one of many reasons why speech intelligibility may be reduced? They need quantitative estimates of speech intelligibility.

The /r/-/l/ contrast among Japanese learners of the English language has been studied by dozens of researchers; many of them reached valuable conclusions. Unfortunately, the value of these efforts for predicting the efficiency of communication is negligible. This thesis reports results on the opposite approach: insight into the fundamentals of speech perception was made subordinate to quantitative description. Now the main question is: was this approach successful in arriving at something of practical value?

An objective speech intelligibility prediction model, the Speech Transmission Index as well as any other, has its own limitations on the scope of the predictions. Extension to the domain of non-native speech communication will not eliminate any other of these limitations, but may potentially even introduce new ones.

Most importantly, the STI model is not capable of predicting the dynamic behavior of people engaged in conversation. For a small part, this behavior can be introduced into the model externally. One may, for instance, choose to include the influence of the Lombard effect (a talker intuitively raising his vocal effort in noise) in STI calculations by assuming a speech level that depends on the ambient noise. Other forms of dynamic behavior

are not as easily modeled. A listener may signal lack of comprehension by his facial expression, prompting a talker to repeat his sentence. Or the talker may respond to an apparent lack of comprehension by avoiding lexically difficult words, reducing his speaking rate, adopting a clearer speaking style or by starting to gesticulate. He may even start speaking louder; whether this is a false reflex based on experience with the hearing impaired or not, it is quite useful in the presence of ambient noise (Chapter 3).

Our approach for prediction of non-native intelligibility is based on the STI method; the effects of these dynamic cues are therefore not included. Unfortunately, communicators appear to rely even more on these cues in non-native scenarios. Intuitive notions of non-native speech intelligibility often *do* include these dynamics: we base our expectations about the non-native speech communication process on our own experiences in practice, which normally include dynamic aspects. This may explain why second-language learners consistently give themselves better ratings for listening than for talking, while measurements show that (if the test sentences are given) non-native production leads to a smaller intelligibility deficit than non-native perception (Chapters 3 and 4). Dynamic behavior is mainly up to the talker, meaning that an L2 learner must work harder when talking than when listening.

Another issue is the performance measure used for rating non-native communication. Our approach is centered around speech intelligibility. For some applications related to non-native speech, intelligibility may not be the most suitable measure. A slight foreign accent, hardly affecting speech intelligibility, may still cause annoyance or trigger prejudice. This can be a reason to pursue perfection of L2 pronunciation beyond the point where foreign accent affects intelligibility.

Also, much of everyday speech communication takes place near 100% sentence intelligibility (the saturation level of the psychometric function). However, a non-native talker may require more attention from his audience to reach 100% intelligibility than a native talker: the cognitive load on the listeners is greater, perhaps even inducing noticeable fatigue.

The list of populations and languages for which the STI correction functions were calculated is limited. This is not really a limitation of the approach; wherever information is missing for a certain application, experiments similar to the ones in Chapters 3 and 4 could be carried out to fill the gaps. However, a specific category of interest has not been addressed: cases where both talker and listener, of the same or different language background, are non-native in the target language. An STI correction function could be derived following the procedure described in Chapter 5, but the situation which it fits will be very specific.

Even given these limitations, a considerable field of applications appears open to the non-native STI-approach described in Chapter 5. A number of possible application examples are outlined below, differentiating between cases where the talker is non-native, and cases where non-native listeners are involved.

Non-native *talkers* can be described in terms of foreign accent strength. A measure of accent strength can be obtained through panels of native judges, but this is relatively time-consuming. Self-ratings can be used as a reasonably accurate, and very quickly and easily obtainable, alternative.

A distribution of self-ratings (grouped into categories) can be converted into STI correction functions such as the ones given in Fig. 5.7. The use of categories instead of a continuous parameter (such as linguistic entropy) imposes limitations on the resolution of the calculations. Since the effects of non-native talking tend to be smaller than non-native listening, a coarser translation into STI correction function appears acceptable. Unfortunately, of several objective acoustic-phonetic measures that were tried out (which could be quickly and easily measured), none was successful as a predictor of speech intelligibility.

Non-native talkers are found in many practical situations. One could think of telephone operators in hotels, tour guides on boat or coach trips, and flight attendants. Estimates of foreign accent strength can be used to impose stricter requirements on the speech transmission quality of equipment (such as cabin sound systems in aircraft), or as a selection criterion when recruiting personnel.

Before being able to obtain intelligibility predictions for a population of non-native *listeners*, rhis population must be described in terms of L2 proficiency. Using the letter guessing procedure, estimates of linguistic entropy can be obtained with relatively limited effort. A distribution of linguistic entropy can be converted to a distribution of the $v$-parameter (and from there, to STI correction functions) using the linear relation shown in Fig. 5.6.

A typical application involving non-native listeners could be the intelligibility of public address systems at international airports. These systems, used to provide information to travelers, can be subjected to speech transmission quality measurements using the STI method, in the usual fashion. Even when messages are repeated in multiple languages, many of the travelers will have to rely on second-language skills. Based on a measured distribution of linguistic entropy among travelers, applicable minimum STI criteria (according to ISO 9921, for instance) can be converted to equivalent minimum-requirements for the (largely non-native) traveler population. This must also involve a deliberate choice: given a certain distribution of (personal) linguistic entropy (LE) among the travelers, to what level do we

wish to adjust the minimum requirements? By using a correction function corresponding to the mean of the LE distribution, we effectively adjust 'native' intelligibility standards to a level where the same intelligibility is reached for 50% of the non-native population. Alternatively, one could decide that at least 90% of the non-native population should understand the messages, and use a correction function corresponding to the 90$^{th}$ percentile of the LE distribution.

Of course, the matter of estimating the distribution of linguistic entropy among travelers is an issue in itself. The fact that the letter guessing procedure is relatively simple and time-efficient opens possibilities that are normally unavailable for experiments in relation to speech intelligibility. One could set up a simple (perhaps even unsupervised) computer terminal in the departure hall, presenting the linguistic entropy test as a game to kill waiting time.

Another example of an application where the listener population may include non-natives, is the field of classroom acoustics. The acoustic properties of a classroom determine if the teacher is sufficiently intelligible to his students. The STI method is very suitable for evaluating classroom acoustics; for classes where some of the students are non-native listeners (immigrants or exchange students), criteria can be adjusted, following a procedure similar to the one describe in the previous example.

Given the abundance of cross-language speech communication in today's world, the need for non-native speech intelligibility assessment tools is likely to become increasingly apparent, for various applications. Some of these will be served by the STI-based approach presented here; for others, this approach will not be as suitable. Even where use of the STI with non-native corrections is not an option, experimental results and models presented in this thesis will hopefully prove to be useful.

# References

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

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

Bergman, M. (**1980**). *Aging and the perception of speech* (University Park Press, Baltimore).

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

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

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 Communication **20**, 255–272.

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

Bronkhorst, A. W., Brand, T. and Wagener, K. (**2002**). "Evaluation of context effexts in sentence recognition," J. Acoust. Soc. Am. **111**, 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. (**1992**), "The intelligibility of English vowels spoken by British and Dutch talkers," in *Intelligibility in Speech Disorders*, edited by R. D. Kent (John Benjamins publishing company, Amsterdam).

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

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

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**, 84–84.

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

Gat, I. N. and Keith, R. W. (**1978**). "An Effect of Linguistic Experience; Auditory Word Discrimination by Native and Nonnative Speakers of English," Audiology **17**, 339–345.

Houtgast, T. and Steeneken, H. J. M. (**1984**). "A multi-lingual evaluation of the Rasti-method for estimating speech intelligibility in auditoria," Acustica **54**, 185–199.

Houtgast, T. and Steeneken, H. J. M. (**2002**), "The roots of the STI approach," in *Past, Present and Future of the Speech Transmission Index*, edited by S. J. van Wijngaarden (TNO Human Factors, Soesterberg).

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 edition "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).

Kanungo, T. and Resnik, P. (**1999**). "The Bible, Truth, and Multilingual OCR Evaluation." In *Proc. of SPIE Conference on Document Recognition and Retrieval (VI),* vol. 3651, San Jose, CA, January 27–28, 1999

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.

Koster, C. J. (**1987**). *Word recognition in foreign and native language; effects of context and assimilation* (Foris, Dordrecht, the Netherlands).

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

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 Teaching **16**, 198–219.

Martens N. and Marciniec, E. (**1977**). "Klein Nederlands-Pools en Pools-Nederlands woordenboek." Wiedza powszechna, Warsaw

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

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

Miller, G. A. and Nicely, P. E. (**1954**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

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**, 2910–2920.

Nábelek, A. K. and Donahue, A. M. (**1984**). "Perception of consonants in reverberation by native and nonnative listeners," J. Acoust. Soc. Am. **75**, 632–634.

Paul, D. B. and Baker, J. M. (**1992**), "The design for the Wall Street Journal-based CSR corpus," in *Proc. Fifth DARPA Speech and Natural Language Workshop* (Morgan Kaufmann Publishers, Inc.), pp. 357–362.

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.

Payton, K. L., Uchanski, R. M. and Braida, L. D. (**1994**). "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," J. Acoust. Soc. Am. **95**, 1581–1592.

Peterson, G. E. and Barney, H. L. (**1952**). "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am. **24**(2), 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.

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

Pols, L. C. W. (**1977**). *Spectral analysis and identification of Dutch vowel in monosyllabic words.* Doctoral thesis, Free University of Amsterdam.

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.

Schouten, M. E. H. (**1975**). *Native-language interference in the perception of second-language vowels.* Doctoral thesis, University of Utrecht.

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

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.

Steeneken, H. J. M. (**1992**). *On measuring and prediction speech intelligibility.* Doctoral thesis, University of Amsterdam

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 Communication **28**, 109–123.

Strange, W. (**1995**), "Cross-Language Studies of Speech Perception; A historical review," in *Speech Perception and Linguistic Experience*, edited by W. Strange (York Press, Baltimore).

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.

Studebaker, G. A. and Sherbecoe, R. L. (**2002**). "Intensity-importance functions for bandlimited monosyllabic words," J. Acoust. Soc. Am. **111**, 1422–1436.

Torgerson, W. S. (**1958**). *Theory and methods of scaling* (Wiley and sons, New York).

van Leeuwen, D. A. and Orr, R. (**2000**), "Speech recognition of non-native speech using native and non-native acoustic models," in *Proc. RTO workshop MIST, RTO-MP-28 AC/323(IST)TP/4*, Neuilly-sur-Seine, France,

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

van Wijngaarden, S. J. (**2000**), "Speech intelligibility of native and non-native speech," in *Proc. RTO workshop MIST, RTO-MP-28 AC/323(IST)TP/4*, Neuilly-sur-Seine, France, pp. 61–66.

van Wijngaarden, S. J. (**2001**). "Speech intelligibility of native and non-native Dutch speech," Speech Communication **35**, 103–113.

van Wijngaarden, S. J. and Houtgast, T. (**2003**). "Effect of speaking styles on the relevance of the Speech Transmission Index in the presence of reverberation," J. Acoust. Soc. Am. **113**, 2296–2297.

van Wijngaarden, S. J. and Steeneken, H. J. M. (**2000**), "The intelligibility of German and English speech to Dutch listeners," in *Proc. ICSLP2000*, Beijing, China, pp. 929–932.

van Wijngaarden, S. J. and Steeneken, H. J. M. (**2001**), "A Proposed Method for Measuring Language Dependency of Narrow Band Voice Coders," in *Proc.. Eurospeech 2001-Scandinavia*, Aalborg, Denmark, pp. 2495–2498.

van Wijngaarden, S. J., Steeneken, H. J. M. and Houtgast, T. (**2001a**), "Methods and models for quantitative assessment of speech intelligibility in cross-language communication," in *Proc. RTO Workshop on Multi-lingual Speech and Language Processing*, Aalborg, Denmark,

van Wijngaarden, S. J., Steeneken, H. J. M. and Houtgast, T. (**2001b**). "The effect of a non-native accent in Dutch on speech intelligibility," J. Acoust. Soc. Am. **109**, 2473.

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**, 3004–3013.

van Wijngaarden, S. J., Steeneken, H. J. M. and Houtgast, T. (**2002b**). "Quantifying the intelligibility of speech in noise for nonnative listeners," J. Acoust. Soc. Am. **111**, 1906–1916.

van Wijngaarden, S. J., Steeneken, H. J. M., Houtgast, T. and Bronkhorst, A. W. (**2002**). "Using the Speech Transmission Index to predict the intelligibility of non-native speech," J. Acoust. Soc. Am. **111**, 2366.

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.

# Spraakverstaanbaarheid in andere talen dan de moedertaal

## Samenvatting

Spraak is een belangrijke vorm van communicatie. De effectiviteit van spraakcommunicatie wordt in hoge mate bepaald door de verstaanbaarheid. Voor veel toepassingen, zoals omroepinstallaties in openbare gebouwen, worden minimumeisen in termen van spraakverstaanbaarheid gesteld. Met objectieve meetmethoden, zoals de Spraak Transmissie Index (STI) kan op basis van relatief eenvoudige metingen de spraakverstaanbaarheid worden voorspeld.

In toenemende mate spelen bij spraakcommunicatie taaloverschrijdende factoren een rol. Hiervan is sprake als één of meer partijen communiceren in een andere taal dan hun moedertaal (vaak wordt de Engeltalige term *non-native* gebruikt). Modellen voor voorspelling van de spraakverstaanbaarheid, zoals de STI, gaan er impliciet van uit dat iedereen communiceert in zijn moedertaal. Om ook voorspellingen te kunnen doen voor toepassingen waar deze aanname niet gerechtvaardigd is, zoals omroepinstallaties op internationale luchthavens, is het noodzakelijk kwantitatieve modellen te construeren voor de invloed van taaloverschrijdende spraakcommunicatie op de verstaanbaarheid.

Om te komen tot een kwantitatief model voor het voorspellen van non-native spraakverstaanbaarheid zijn verstaanbaarheidsexperimenten uitgevoerd. Hierbij is onderscheid te maken tussen experimenten waarbij de spreker, en experimenten waarbij de luisteraar de "non-native factor" is.

Het spreken van een vreemde taal gaat vrijwel altijd gepaard met een accent. Dit accent belemmert de verstaanbaarheid; de sterkte van het accent (beoordeeld door *native* luisteraars, met behulp van paarsgewijze vergelijking) correleert sterk met de verstaanbaarheid van zinnen. Tevens blijkt de eigen mening van non-native sprekers over de sterkte van hun accent (uitgedrukt in een waardering tussen 1 en 5) een acceptabele voorspeller voor het effect van het accent op de spraakverstaanbaarheid. Tevens zijn de mogelijkheden onderzocht om door middel van akoestisch-fonetische metingen aan het spraaksignaal objectieve maten te bepalen die de verstaanbaarheidseffecten van een accent voorspellen. Helaas blijken de beoordeelde maten (uitgezocht

op basis van gemak en snelheid waarmee ze te bepalen zijn, en praktische toepasbaarheid) niet succesvol.

Wanneer verstaanbaarheid door middel van proefpersoonexperimenten wordt bepaald, wordt vaak gebruikt gemaakt van verstaanbaarheidsmaten die zijn gebaseerd op de correcte herkenning van woorden en/of fonemen. Foneemherkenning en woordherkenning blijken voor non-native spraak slechte voorspellingen voor verstaanbaarheid van zinnen op te leveren. Non-native sprekers hebben vaak een incorrect begrip van foneemcategorieën, en maken allerhande fouten in de uitspraak. Dit heeft uiteraard een negatief effect op de verstaanbaarheid; echter, door gebruik te maken van door zinscontext geïntroduceerde redundantie kunnen (native) luisteraars veel van de verloren gegane informatie reconstrueren, waardoor de gevolgen voor de verstaanbaarheid beperkt blijven. Verder worden ook uitspraakfouten gemaakt die op woordniveau slecht waarneembaar zijn, zoals prosodische onvolkomenheden. Omdat in de praktijk vooral de verstaanbaarheid van gehele boodschappen van belang is, zijn woord- en foneemherkenning niet geschikt voor het beoordelen van de verstaanbaarheid van non-native spraak.

Bij luisteraars blijken de taalkundige vaardigheden in de vreemde taal waarnaar wordt geluisterd van groot belang. Deze vaardigheden kunnen worden beoordeeld door de linguïstische entropie van luisteraars te bepalen, met behulp van een taak waarbij men opeenvolgende letters in een zin moet raden. Linguïstische entropie is een goede voorspeller van verstaanbaarheid bij non-native luisteraars, waaruit het belang van de taalkundige context voor de verstaanbaarheid blijkt. Dit heeft tevens praktische betekenis: linguïstische entropie is aanzienlijk sneller en gemakkelijker te meten dan spraakverstaanbaarheid.

Gegeven een zeker taalvaardigheidsniveau zijn de gemeten verstaanbaarheidseffecten bij het spreken groter dan de effecten bij het luisteren. Dit is strijdig met het intuïtieve gevoel van velen, dat onder andere tot uitdrukking komt doordat men zichzelf in vreemde talen betere luister- dan spreekvaardigheden toedicht. Dit is te verklaren doordat in conversaties een slechte verstaanbaarheid ertoe leidt dat vooral de spreker zich moet aanpassen – bij het spreken van een vreemde taal moet harder gewerkt worden dan bij het luisteren.

Om objectieve voorspellingen van de verstaanbaarheid (de Spraak Transmissie Index) te verkrijgen in taaloverschrijdende scenario's, is het noodzakelijk om het STI-model aan te passen. Er is gekozen voor een aanpak waarbij de berekening van de STI niet wordt gewijzigd; metingen kunnen derhalve met bestaande apparatuur worden uitgevoerd. In plaats daarvan wordt de *interpretatie* van de STI afhankelijk gemaakt van de (non-native) populatie van sprekers en luisteraars.

De STI wordt normaal gesproken vertaald naar verstaanbaarheids-kwalificaties door een gestandaardiseerde tabel met semantische labels ('slecht'-'uitstekend'), met bijbehorende grenzen in termen van STI. Deze tabel wordt door middel van een correctiefunctie aangepast voor non-native populaties; dezelfde labels worden van andere STI-grenzen voorzien.

Voor het berekenen van STI-correctiefuncties wordt gebruik gemaakt van psychometrische functies (zinsverstaanbaarheid als functie van de spraak-ruisverhouding). Door vergelijking tussen native en non-native psychometrische functies kan voor elk verstaanbaarheidsniveau (percentage zinsverstaanbaarheid) het verschil in benodigde spraak-ruisverhouding tussen native en non-native spraak worden berekend. Dit verschil in spraak-ruisverhouding wordt vervolgens vertaald naar een verschil in STI, waaruit de benodigde correctie volgt.

Om de correctiefunctie te kunnen baseren op een minimaal aantal modelparameters, wordt een nieuw model voor de non-native psychometrische functie geïntroduceerd: de non-native psychometrische functie wordt gerelateerd aan de native psychometrische functie door middel van één parameter ($\nu$). De native psychometrisch functie wordt, zoals vaak het geval is, verondersteld zich als een cumulatieve normaalverdeling te gedragen (bepaald door gemiddelde en standaarddeviatie, $\mu$ en $\sigma$). De STI-correctiefunctie kent derhalve drie parameters, waarvan er slechts één ($\nu$) apart hoeft te worden bepaald voor elke non-native populatie.

Voor luisteraars kan de parameter $\nu$ door middel van een experimenteel afgeleide lineaire relatie worden berekend uit schattingen van de linguïstische entropie. Wanneer deze voor een populatie sprekers of luisteraars bekend is, kan derhalve een correctiefunctie worden berekend, waarmee de tabel met kwalificatie-labels wordt aangepast voor de betreffende populatie.

Sprekers kunnen, op basis van paarsgewijze beoordeling of op basis hun eigen mening met betrekking tot hun accent, worden ingedeeld volgens een systeem van categorieën. Met elke categorie correspondeert een waarde van $\nu$, waarmee de STI-kwalificatietabel wordt aangepast.

De gehanteerde aanpak voor het bepalen van correctiefuncties vereist dat het STI-model alle vormen van spraakdegradatie (zoals ruis, nagalm, echo's en oversturing) gelijkwaardig in rekening brengt. Wanneer de STI wordt gebruikt voor het voorspellen van zinsverstaanbaarheid, blijkt in bepaalde gevallen het effect van nagalm op de verstaanbaarheid te worden onderschat. Dit blijkt samen te hangen met de gehanteerde spreekstijl. Door een aanpassing in het bereik van modulatiefrequenties waarop de STI-berekening is gebaseerd, wordt voor een informele, conversationele spreekstijl de nauwkeurigheid van de STI-methode verbeterd.

Bij de experimenten is voorbijgegaan aan de dynamische, adaptieve aspecten van spraakcommunicatie: luisteraars geven (bijvoorbeeld door interrupties, of door hun gezichtsuitdrukking) aan wanneer de verstaanbaarheid te wensen overlaat. Sprekers reageren hierop, bijvoorbeeld door duidelijker en langzamer te spreken. Dergelijke effecten zijn niet in de voorspellingsaanpak meegenomen. Deze aanpak is dan ook vooral te gebruiken voor situaties die geen gelegenheid tot dergelijk adaptief gedrag bieden, zoals het gebruik van omroepinstallaties, en situaties waarbij grote groepen tegelijk worden toegesproken.

# Appendix A. Derivation of an STI correction function based on a logistic function

Deriving a correction function based on the psychometric functions described by Eqs. 5.2 and 5.4, 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 $\Phi\left(\frac{r-\mu}{\sigma}\right)$ may be approximated by a logistic function (e.g., Versfeld et al., 2000), such as Eq. A2

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

where

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

By substituting $\Lambda(\rho)$ for $\Phi\left(\frac{r-\mu}{\sigma}\right)$ in Eq. A1 and solving, Eq. A4 is obtained.

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

By substituting Eqs. 5.1 and A3 in Eq. 5.4, the correction function Eq. A5 is obtained.

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

# Curriculum Vitae

Sander van Wijngaarden werd geboren op 1 oktober 1971 in Rotterdam. Hij groeide op in Rotterdam, maar doorliep de middelbare school in Capelle aan den IJssel (Christelijke Scholengemeenschap Comenius). Na het behalen van zijn VWO-diploma in 1990 begon hij met de studie Technische Natuurkunde aan de Technische Universiteit Delft. Naast studeren hield hij zich bezig met allerlei nevenactiviteiten, zoals student-lidmaatschap van faculteitsraad en universiteitsraad, voorzitterschap van de Delftse Studentenraad (de ledenraad van de studentenvakbond) en lidmaatschap van het eettafelbestuur van DSV Sint Jansbrug. Zijn afstudeeronderzoek ("het toonhoogtediscriminatie-vermogen bij monotische en dichotische kamruis") verrichtte hij bij de sectie Akoestische Perceptie. Kort na zijn afstuderen trad hij in 1996 in dienst van TNO Technische Menskunde in Soesterberg, als wetenschappelijk medewerker in de groep Spraak. In deze functie hield hij zich in aanvang met name bezig met projecten op het gebied van de menselijke spraakperceptie, gehoorbescherming en communicatie. Na in 1999 een beperkte studie op het gebied van *non-native* spraak te hebben uitgevoerd, begon hij in 2000 onder begeleiding van prof.dr.ir. T. Houtgast en dr.ing. H.J.M. Steeneken aan een driejarig achtergrond-onderzoeksproject, waaruit deze dissertatie is voortgekomen. Sinds 1 april 2003 is hij onderwerpcoördinator van de groep Spraak & Gehoor.

The ball flew over the fence Míč přeletěl přes plot Bollen flög över staketet 球飞过了篱笆

Kamuolys perlėkė per tvorą Мяч перелетел через забор Lopta je preletjela preko ograde Топката п

плот Bola telah terbang melintasi pagar 球は塀を越えた A bola fou por riba da cerca Bolden

føyk over gjerdet A bola que passou a voar por cima da cerca De bal fleach oer 't sket Mingea a zb

прелете преко ограде Chuaigh an liathróid thar an sconsa Η μπάλα πέταξε πάνω από τον φράχτη

球飞过了篱笆 Quả bóng bay qua hàng rào Pallo lensi aidan yli A labda átrepült a kerítés fel

preko ograde Топката прелетя над оградата הכדור עף מעבר לגדר La palla volò via oltre il recinto

por riba da cerca Bolden fløj over stakittet Der Ball flog über den Zaun Lopta preletela cez plot La

oer 't sket Mingea a zburat peste gard Weẽñ we bal õuch waa Baloia hesi gainetik ihes egin zaig

πάνω από τον φράχτη توپ از روی دیوار پرت شد The ball flew over the fence Míč p

labda átrepült a kerítés felett La pelota voló por encima de la valla Kamuolys perlėkė per tvorą Мяч

volò via oltre il recinto الكرة طارت فوق الجدار Piłka przeleciała przez płot Bola telah terbang melintasi

preletela cez plot La balle passait au-dessus de la clôture Ballen føyk over gjerdet A bola que passo

gainetik ihes egin zaigu Top duvarın üzerinden uçmuştu Лопта прелете преко ограде Chuaigh an

over the fence Míč přeletěl přes plot Bollen flög över staketet 球飞过了篱笆 Quả bóng b

perlėkė per tvorą Мяч перелетел через забор Lopta je preletjela preko ograde Топката прелетя на

telah terbang melintasi pagar 球は塀を越えた A bola fou por riba da cerca Bolden fløj over st

gjerdet A bola que passou a voar por cima da cerca De bal fleach oer 't sket Mingea a zburat peste

преко ограде Chuaigh an liathróid thar an sconsa Η μπάλα πέταξε πάνω από τον φράχτη شد

球飞过了篱笆 Quả bóng bay qua hàng rào Pallo lensi aidan yli A labda átrepült a kerítés fel

preko ograde Топката прелетя над оградата הכדור עף מעבר לגדר La palla volò via oltre il recinto

por riba da cerca Bolden fløj over stakittet Der Ball flog über den Zaun Lopta preletela cez plot La

oer 't sket Mingea a zburat peste gard Weẽñ we bal õuch waa Baloia hesi gainetik ihes egin zaig

πάνω από τον φράχτη توپ از روی دیوار پرت شد The ball flew over the fence Míč