# English as a lingua franca: mutual intelligibility of Chinese, Dutch and American speakers of English

Wang, H.

# Chapter two

# Background

## 2.1 Foreign accent

Languages differ, and people from different places speak differently. Everyone may have had the experience, when listening to a foreigner speaking his/her own language, of having great difficulty in understanding what he is trying to say, not because of the speaker's lack of knowledge of vocabulary and language structure but because the sounds he produced seemed peculiar and because his voice rose and fell in unexpected places.

With the development of globalization and internationalization there is more and more communication involving speakers from many different linguistic and cultural backgrounds. Internet and cheap intercontinental telephony make oral communication feasible between people from anywhere in the world. Internet conferencing would be an ideal way for researchers to exchange ideas and to save time, money and energy as well, if they could really talk to each other without problems. Unfortunately, on many occasions, communication breaks down because the listener cannot get a clear idea of what his interlocutor is trying to say, due to his deviant pronunciation and speech melody. The consequences of such non-native communication may be severe if it happens in the air traffic control tower, or hospital emergency room, when people from different language backgrounds who need urgent information or help, cannot make themselves understood.

### 2.1.1 What is a (foreign) accent?

As a distinctive manner of oral expression, the notion of accent has two uses in linguistics. On the one hand accent refers to the way a speaker uses to make a syllable stand out in a word (word stress) or to make a word stand out in a constituent or sentence (sentence stress) so as to mark the syllable or word as communicatively important in the spoken utterance. To this effect the speaker may employ a variety of phonetic means, such as more careful pronunciation, greater loudness, longer duration and a relatively sudden change in vocal pitch (see for instance Van Heuven and Sluijter, 1996; Nooteboom, 1997). On the other hand, accent may refer to the way of speaking that is characteristic of a specific group of people from a regional background. What both readings of the term accent have in common is that some entity, be it a syllable, a word, or a speaker, stands out from its background. This thesis is about the second meaning of accent, i.e. deviant pronunciation rather than prosodic prominence.

People in different regions speak differently even in the same country in the same language. A regional variety of a language differing from the standard language is called a **dialect** when it is distinguished by differences at several linguistic levels, e.g. in pronunciation, grammar and vocabulary. When there are no differences in grammar and vocabulary but only the pronunciation (including the rhythm and melody) differs, the language variety is called an **accent** or **local/regional accent (L1 accent)**. Everybody speaks with some sort of an accent as a pattern of speech production. It betrays the speaker's geographical background, socio-economic class, ethnic identity, educational level, etc. Normally, the more distant the speaker's region is from that of the listener, the more different the accents of the interlocutors are, and the more difficult it is for them to understand each other.

When people learn a foreign language (L2), especially after puberty, they do not normally acquire native pronunciation in the new language. They will typically speak the foreign language with an accent, which is often the result of substituting phonemes and/or allophones of the native language (L1) for sounds that are needed in the foreign language. This kind of accent is called **foreign accent (L2 accent)**. Broadly speaking, then, foreign-accented speech is non-pathological speech produced by second-language users that sounds noticeably different from the speech of native speakers of the target language. It is probably true that there is little or no principled difference between speaking a language with a regional (native) accent or with a foreign accent. In both cases structures from the native dialect or language are transferred to the target language – be it the standard variety of one's native language or to a foreign language.

### 2.1.2 Linguistic levels in foreign accent

Even though we have restricted the notion of (foreign) accent to non-native language varieties that differ from the native norm only in terms of the sounds, it is not unusual to subdivide this area into the more abstract, representational aspects called **phonology** versus the more concrete aspects of the implementation of the abstract categories which are subsumed under the heading of **phonetics**. Phonologically foreign accent is often seen as wrong / missing representations of phonemes in the second language; phonetically, foreign accent is primarily the incorrect phonetic output routine which is employed to implement a correct phonological representation. Phonetic deviance is readily detectable by native listeners and can arise from phonemic, subphonemic, or suprasegmental differences in speech production (Flege, 1995). A phonemic difference would be the failure to distinguish between two members of a contrast in the target language because there is no such contrast in the learner's mother tongue, cf. Chinese (and Dutch) learners of English do not have distinct sound categories for the phonemes /ɛ/ and /æ/. An example of a subphonemic difference would be the failure to observe certain positional allophonic variants of a phoneme, such as the use in English of clear /l/ in the onset versus dark /l/ in the coda, when the learner's native language does not feature this allophonic difference, as would be the case for a French learner of English. A suprasegmental difference at the level of phonetics would be, for instance, the way Japanese learners of English would fail to mark English stressed syllables by greater duration and loudness, as their native language marks stress by pitch only (Beckman 1986).

### 2.1.3 Relative importance of pronunciation, morpho-syntax and vocabulary for intelligibility and comprehensibility

The question that we raise in this section is whether foreign-accented speech is indeed more difficult to understand than native speech. It should be stated at the outset that communication between a foreign speaker and a native listener is generally unproblematic as long as the foreign accent is relatively mild and the communication channel is noiseless.[1] For instance, Munro and Derwing (1995a) showed that the word error rate of their Mandarin learners of English was 11% against 4% for native English control speakers.[2] It is not easy to interpret such a finding. On the one hand the intelligibility of the foreign speakers is still quite high, since nine out of every ten words are correctly recognized. On the other hand, the error rate of the Mandarin learners is three times as high as that of the native speakers. Munro and Derwing used studio-quality recordings played back to listeners under high-fidelity conditions, unrealistic of real-life communication. Also, the Mandarin learners of English were immigrants to Canada with a minimum length or residence in excess of one year. The quality of their pronunciation must have been a lot better than that of the more typical Chinese speaker of English without any experience in an English-speaking environment.

Under more averse communicative circumstances, predictably, the intelligibility of foreign-accented speech deteriorates relative to native speech. As a case in point, Van Wijngaarden (2001) showed the effect of native versus non-native speech by adding noise to the communication channel. He defined an intelligibility threshold ('Speech Reception Threshold' or SRT) at 50% correct sentence recognition. His results showed that intelligibility was at threshold at a –6dB speech-to-noise ratio between native L1 Dutch speakers and listeners. When the speakers were English learners of Dutch, communication was less robust, at an SNR of –2dB, indicating that non-native speech is clearly less resistant to noise. This ties in with the subjective opinions of native listeners when exposed to samples for foreign-accented and native speech. The former type is uniformly judged to more difficult to understand. It seems to be the case, then, that when judging the difficulty of accented-speech; the judges have a clear conception of how well speech samples will hold up under averse listening circumstances.

---

[1] Lane (1963) seems to have been the first to establish that word recognition by native listeners is poorer for foreign-accented than for native-accented speech utterances. He found that word recognition for Serbian-, Japanese- and Punjabi-accented English was approximately 36% poorer than for native-English speech in a range of signal-to-noise ratios and filtering conditions. Lane's results, then, also indicated that the effect of foreign accent was greatly reduced as the speech channel was relatively noiseless.

[2] It is not easy to compute the word error rates from the data presented by Munro and Derwing (1995a). Given a mean utterance length of 10.7 words and three utterances contributed by each of ten Mandarin learners and two native control speakers, which were orthographically transcribed by 18 native listeners I divided the total number of word errors obtained for the Mandarin learners (636) by 5,778 and that of the control speakers (44) by 1,156.

Foreign learners of a target language deviate from the native norm not only in terms of pronunciation but also in their use of words and morpho-syntactic structure (so that it would be more apt to speak of foreign dialect, see above). We might therefore ask the question whether getting the pronunciation right should be a greater or lesser concern for the foreign learner than getting the lexis and morpho-syntax right. For several decades, pronunciation experts have stressed improving intelligibility as the most important goal of pronunciation teaching. As early as 1949, Abercrombie argued that most "language learners need no more than comfortably intelligible pronunciation" (p.120). This view has been echoed more recently by Gilbert (1980), Pennington and Richards (1986), Crawford (1987) and Morley (1991). However, this does not necessarily mean that improving one's pronunciation is the only – or even the most important – way to become a more intelligible speaker of a foreign language.

Several researchers have attempted to isolate the role of pronunciation, as compared to other linguistic features, in speech understanding. Gynan (1985) found that listeners judged that the phonology of Spanish non-native speakers of English interfered with their comprehensibility to a greater extent than grammatical errors did. Ensz (1982), on the other hand, found that grammar was more important than pronunciation for speech understanding when American non-native speakers were judged by native speakers of French. In a study of English-accented German, Politzer (1978) found that vocabulary errors affected listening comprehension most significantly, followed by grammar and then by pronunciation. In the study by Munro and Derwing (1995a) discussed above, the authors correlated the number of pronunciation errors and syntactic errors with objective (word error rate) and subjective (opinion scores) intelligibility measures. Grammatical errors correlated more strongly than phonemic errors with subjective intelligibility whilst the reverse was true for the objective word error measure. Later the same year Munro and Derwing (1995b) tested comprehension (by a sentence verification task) and processing time (for correct verification only) of 20 native English listeners who were exposed to the production of 10 Mandarin-accented speakers of English and 10 native English control speakers. Ninety-nine percent correct verification was obtained for the control speakers against 93% for the Mandarin speakers. Moreover, correct verification took about 60 ms longer for the Mandarin-accented utterances than for the control utterances. The native English utterances received much better accent ratings (mean = 1.5 on a scale from 1 to 9) than the Mandarin-accented counterparts (mean = 6.3). The same was true for subjective comprehensibility ratings (1.5 versus 5.4 for native versus foreign-accented tokens). In both studies (Munro and Derwing 1995a, b), judged comprehensibility and accentedness correlated around $r = 0.624$, which correlation is similar to the $r = 0.580$ reported by Van Heuven, Kruyt and De Vries (1981) but considerably less than the $r = 0.889$ that was found by Varonis and Gass (1982).

It should be noted that the articles reviewed here studied the relative strength of pronunciation versus morpho-syntactic errors as determinants of intelligibility through a correlational approach. To the best of my knowledge only Van Heuven (1986) varied morpho-syntactic and phonemic errors in an orthogonal experimental design. In his study of the intelligibility of native versus Turkish-accented Dutch, he varied the quality of the pronunciation (native, foreign) independently of the

morpho-syntactic properties (native, foreign) and found that the effects of pronunciation were roughly twice as large as those of morpho-syntactic deviations.[3] Native Dutch pronunciation resulted in 23% more correctly understood utterances (and 145 ms faster reaction times) than the Turkish-accented counterparts. The effects of Dutch versus Turkish morpho-syntax were a difference of 12 % and 93 ms, about half as large as the effect of pronunciation.

One could also argue on logical grounds that a good-quality pronunciation in a foreign language has higher priority than proper grammatical and morphological structure. Generally, a speaker can make himself understood in a foreign language as long as the content words are intelligible; the exact order in which the morphemes and words reach the listener would seem to be of secondary importance. After all, for word-order to have a (positive or negative) effect on intelligibility, the listener should first recognize the words: without any words there would be no word order to begin with.

### 2.1.4 Relative importance of various aspects of pronunciation (vowels, consonants, stress, accentuation, melody, rhythm)

A number of findings in foreign-accented speech research have emerged over the years with respect to those characteristics of speakers that were associated with either a greater or lesser degree of perceived foreign accent. Specific characteristics of the tokens produced by speakers have been associated, in various studies, with degrees of perceived foreign accent. Little is known about the relative importance of errors at each of the various linguistic levels on intelligibility of foreign-accented speech and perceived strength of foreign accent. Moreover, it may well be the case that particular errors are highly conspicuous and yet do not interfere with intelligibility, whereas other errors may go more or less unnoticed but are quite harmful to intelligibility. As a case in point, it has often been found that deviations in vowel quality and duration are very noticeable in foreign-accented speech. Yet, native listeners of languages such as English and Dutch are extremely flexible when they process utterances with incorrectly pronounced vowels. Van Ooijen (1994) showed that when confronted with nonwords that differed from their nearest lexical word in either one vowel or one consonant, listeners were much quicker to correct the vowel than the consonant. It has been argued (e.g. Best, 1993) that errors in vowels, which have greater intensity and duration than consonants, should be more detrimental to intelligibility than consonant errors.

There are indications that incorrect placement of word stress in English is highly detrimental to intelligibility. In good-quality native speech stress errors are not a problem, but stress is very important in speech of poor segmental quality, such as computer speech, speech in noise, and foreign-accented speech. It would appear that the stress pattern serves to limit the lexical search space for the native listener. When the stress pattern is incorrect, the listener will reinterpret the segments so that a word is found within the incorrectly constrained sublexicon. Examples that speak to the issue are given by Bansal (1966), quoted in Cutler (1983: 79), for Indian

[3] Van Heuven (1986) is a summary in English of earlier work reported in more detail in Dutch by Van Heuven, Kruyt and De Vries (1981) and Van Heuven and De Vries (1981, 1983).

English. In the Indian pronunciation of English the stress is perceived by English listeners one syllable later than where the Indian speaker intends it to be. As a result *character* was perceived as *director* and *written* as *retain*. However, it would seem to me that such effects will be restricted to languages that have contrastive stress. If a language has fixed stress or no stress at all, deviations from the canonical stress pattern will not greatly interfere with speech intelligibility.

It has often been said that speech melody has little impact on speech intelligibility.[4] The relative unimportance of melody is also suggested by the practice of state-of-the-art speech recognition software. There is not a single automatic speech recognizer that uses melodic information; the words can be recognized quite well just by identifying their constituent vowels and consonants. Nevertheless, Van Wijngaarden (2001) showed that the intelligibility of electronically monotonized Dutch speech (as defined by the Speech Reception Threshold, i.e. the signal-to-noise ratio at which 50% word recognition was still possible) was more difficult than the same utterances with melody intact (a 2-dB change in SRT).[5]

When dealing with speech melody, one has to make a clear distinction between melody at the level of the sentence (as in the preceding paragraph) and at the word level. It would seem obvious that incorrect word tones would greatly reduce the intelligibility of monotonous speech in tone languages, especially when the language has a predilection for short, monosyllabic words with a simple CV structure and has a large inventory of lexical tones – such as Chinese languages (with at least four tones, as in Mandarin, up to ten or more as in Cantonese). We are not aware of any studies of the intelligibility of monotonized speech in tone languages.

The upshot of the above is that it is very difficult to make generalizations as to the relative importance of specific levels in the linguistic hierarchy for intelligibility. Too much depends on the structural differences between source and target languages at each of the levels; therefore, what seems a clear difference of one level in favour of another in one language pair may be reversed in another pair.

More detailed studies have addressed the relative importance of specific types of error for the detection of foreign accent. Magen (1998) edited Spanish-accented English phrases so as to correct elements thought to be associated with the foreign accent. Adjustments to syllable structure, consonant manner of articulation and word stress were found to produce the most substantial effects in decreasing degree of perceived foreign accent. Adjustments to voice onset time (VOT), on the other hand, had little effect. Gonzalez-Bueno (1997) considered the role of stop voicing by manipulating the voice onset time of the initial segment /k/ in the Spanish word *casa* 'house' spoken by a native speaker of English. In judging the foreign accentedness of the single word token, raters identified those instances where the VOT of the /k/

[4] Obviously, speech melody is much more important for speech understanding. The chunking of the stream of speech in phrases and the highlighting of important words within the phrases, as well as the signalling of clause type, depend on the intonation pattern. Since the present thesis is about speech intelligibility rather than understanding, this role of intonation will not be considered.

[5] Here Dutch listeners heard native Dutch speakers. The effect of monotonization was a 3-dB poorer SRT when English-accented speech was presented to Dutch listeners.

was between 15 and 35 ms as most native-like, suggesting that VOT may indeed influence the degree of perceived accentedness. Using natural stimuli collected in a longitudinal study of English pronunciation by Japanese learners, on the other hand, Riney and Takagi (1999) only found limited support for a correlation between the VOT of stop segments and global foreign-accent ratings.

The accuracy of liquid pronunciation has also been considered within this context, though again, the relationship between segmental accuracy and global accent has not been firmly established

Major (1986) found that among native speakers of Brazilian Portuguese learning English, higher rates of epenthesis were significantly correlated with stronger global foreign accent. The use of epenthetic /i/ as opposed to schwa was particularly indicative of stronger accent.

Prosodic aspects of speech have also been demonstrated to correlate with global foreign accent. Magen (1998) and Major (1986) found that when all segmental information was removed from the speech stream judges were able to distinguish between English passages spoken by native speakers of English and native speakers of Mandarin. Jilka (2000) found similar results, with the accuracy of sentence level intonation being significantly correlated with the degree of perceived foreign accent in German speech of native speakers of English. Anderson-Hsieh, Johnson and Koehler (1992) echo the importance of prosodic factors in influencing perceived foreign accent, identifying them as more important than segmental and syllable structure factors in their study of English learners from range of L1 backgrounds.

The influence of speech rate has also been considered by some researchers. MacKay, Meador and Flege (2001) found that among late Italian-English bilinguals shorter sentences were perceived to be less foreign accented. Munro and Derwing (1998) found that the English speech of native speakers' of Mandarin was deemed to be more accented when slowed down and that at least some speakers' accents were found to be less strong when their speech was speeded up. Munro and Derwing (2001) suggest that the natural speaking rate of non-native speakers is typically somewhat slower than optimal (i.e. native) and found that when foreign-accented speech needed only slight speeding-up in order to be perceived as less foreign.

Interestingly, however, comprehensibility and intelligibility have been found to be only moderately correlated with global foreign accent scores (Munro and Derwing 1995a, 1999) in the English speech of native speakers of Mandarin. Munro and Derwing (1995b) pointed out that even highly accented speech can still be intelligible and comprehensible to native speakers.

### 2.1.5 Attitudes towards foreign accent

We have seen in the preceding section that, although the effects of foreign accent may be relatively small in terms of intelligibility and comprehensibility of speech utterances communicated through a virtually noiseless channel, native listeners seem to hear immediately that a speaker has an accent. Given, then, that foreign accent is readily detectable even when it does not overtly influence intelligibility, we may ask if native listeners are annoyed by foreign-accented speech or even discriminate against speakers with a foreign accent. Indeed, there has been a long tradition of research on attitudes towards foreign accent, as one of the salient characteristics of

L2 learners. A wide range of studies has shown that listeners often evaluate foreign accented speech negatively (Brennan and Brennan, 1981a, b; Fayer and Krasinski, 1987; Kalin and Rayko, 1978; Ryan and Carranza, 1975).

Aristotle (384-322 B.C.E.) believed that the type of language which speakers use has an effect upon their credibility or *ethos* (trans. Cooper, 1932). A similar idea is apparent in the Renaissance rhetoricians' preoccupation with the details of verbal expression. Research by dialect geographers in the early twentieth century called attention to language varieties which were stigmatized or, on the other hand, accorded prestige (Bloomfield, 1933). The earliest contemporary research on language attitudes towards language varieties was done by Lambert, Hodgson, Gardner and Fillenbaum (1960). In the 1970's, researchers continued to study attitudinal consequences of ethnically and regionally determined language variation. These numerous studies have shown that native listeners tend to downgrade non-native speakers simply because of foreign accent (Lambert, Hodgson, Gardner and Fillenbaum, 1960; Anisfeld, Bogo and Lambert, 1962; Ryan and Carranza, 1975; Kalin and Rayko, 1978; Brennan and Brennan, 1981a, b).

A significant body of research shows that foreign-accented speakers may be viewed as less intelligent, less competent, and even less attractive than native speakers. Rubin and Smith (1990) conducted a matched-guise study in which two Chinese women produced mini-lectures in both moderately and highly accented English. Intelligibility was functionally tested using a Cloze Blank-filling test. In subsequent opinion tests among various dimensions, one was to rate the instructor in terms of accentedness and teaching ability. Crucially, during the listening session half of the students saw a picture of a Caucasian woman and half saw a picture of a Chinese woman. The results showed that the students did not distinguish between the highly and moderately accented conditions but were affected by the suggested ethnicity of the speaker. Objective intelligibility scores and perceived accentedness were poorer when the Asian speaker was suggested; moreover, the impression of the instructor's teaching ability was negatively correlated with perceived accentedness. It seems to me that the listeners held a stereotypical expectation of the Chinese speaker being poorly intelligible (quite probably based on real-life experience), which caused them to make a less motivated effort to understand the speaker, i.e. the listeners gave up the attempt even before they had really tried.

Schinke-Llano (1983, 1986) noted that classroom teachers are often reluctant to engage English L2 students in conversation beyond basic classroom management exchanges. All these findings suggest that early intelligibility problems with foreign-accented speakers may have negative attitudinal and communicative effects on later exchanges with similarly accented speakers.

The negative effects of foreign accent have been found to extend beyond the classroom. Some evidence indicates that people in English-speaking regions in Canada have been denied housing or employment simply because of a French accent (from Munro's website).[6] The discrimination of foreign accent appears to have catalyzed the rise of accent-reduction programs which aim to reduce or eliminate foreign accents altogether.

---

[6] http://www.sfu.ca/%7Emjmunro/research.htm.

It would be wrong to conclude from the evidence above that foreign accent is always a social handicap. Research in the Netherlands (Doeleman, 1998) presents a more balanced view. Some foreign accents were found to be prestigious (Dutch spoken with an American, or better still, British accent) whereas other accents were attributed low prestige (e.g. Surinam, Moroccan and Turkish accent). The status of the accent seemed tied to the status of the community whose language is the source of the accent.

## 2.2 Causes of foreign accent

Now that we have defined what we mean by foreign accent and have briefly considered (some of) its communicative effects, let us consider factors that may cause foreign accent. Given that foreign accent in speech production is tantamount to saying that the sounds produced by the learner are off-target, we may ask what factors limit the phonetic accuracy in foreign language speech production. Moreover, it has often been noted that some learners have a stronger, more noticeable foreign accent than others. What, then, makes one L2 speaker have a more or less heavy accent than another? What factors contribute (most) to cross-language variation in foreign accent?

### 2.2.1 Age effects (AOA and AOL)

Age of arrival (AOA) and age of learning (AOL) are important factors for foreign-accented speech. AOA refers to the first arrival time of the L2 learner in a predominantly target language speaking country. AOL refers to the chronological age at which an individual first begins receiving massive input from native speakers of an L2 in a naturalistic context. Although very young immigrant children may arrive in the new country a few years earlier than they are exposed to the L2 (typically not before they go to school), AOA and AOL generally coincide. We will therefore no longer distinguish between them.

Taking a cue from Lorenz' (1961) work on imprinting in ducks and geese, Lenneberg (1967) introduced the critical period concept to research in native-language acquisition and claimed that foreign accent in an L2 cannot be overcome easily after puberty, because after puberty the ability for self-organization and adjustment to the physiological demands of verbal behavior quickly declines.[7] The brain behaves as if it has become set in its ways and primary, basic skills not acquired by that time usually remain deficient for life. Flege, Yeni-Komshian and Liu (1999) suggest that age affects phonology more than morphosyntax.

Many researchers support the view that age of learning is a very significant determinant of the degree of foreign accent. Long (1990) concluded that the L2 is

[7] Originally the phrase 'critical period' was used in ethologists' studies of species-specific behavior. It is the period when imprinting is observed in certain species such as young birds and rats. For example, geese isolated from their parent birds since the hatching react to and follow the moving object they see first. This kind of behavior can be learnt only during a short period of time after hatching (Lorenz, 1961, quoted in Clark and Clark, 1977).

generally spoken without an accent up to an AOL of 6 years, with a foreign accent by nearly all subjects having AOLs greater than 12 years, and either with or without foreign accent by subjects in the intermediate AOL range. Flege and Fletcher (1992) provided indirect evidence that foreign accent may be evident in the speech of adults who began learning their L2 as early as 7 years of age. As far as the pronunciation of an L2 is concerned, many studies have shown that earlier is usually better, i.e., people who arrive in a target language community at an early age have an advantage over those who arrive as adults (Asher and Garcia, 1969; Selinger, Krashen and Ladefoged, 1975; Oyama, 1976; Suter, 1976; Purcell and Suter, 1980; Tahta, Wood and Lowenthal, 1981a, b; Flege, 1988; Patkowski, 1990; Thompson, 1991; Flege and Fletcher, 1992; Flege, Yeni-Komshian and Liu, 1999). Both the proportion of individuals observed to speak their L2 with a detectable accent, and the strength of perceived foreign accent among individuals with detectable foreign accent have been found to increase as the age of learning the L2 increased. Results of Flege and co-workers show that in the production of several English consonants, Italian bilinguals whose AOL was earlier than 11 years generally performed better than those whose AOL was later than 21 years (Flege and Fletcher, 1992; Flege, Munro and Mackay, 1995; Piske, Mackay and Flege, 2001). These researchers proposed that even when other variables such as length of residence are partialed out, age of learning remains the most critical predictor of degree of foreign accent.[8]

### 2.2.2 Experience effects (LOR and L2 USE)

Two more factors that often come up as potential determinants of degree of accent are Length of Residence (LOR) and intensity of L2 use (USE). LOR is defined as the number of years spent by the learner in a country where the L2 is the predominant language. USE refers to how much/how often the learners use their L2 in daily life. Researchers have largely failed to reach agreement on the existence of a significant correlation between the accuracy of L2 pronunciation and either LOR or USE (Oyama, 1976; Flege and Fletcher, 1992; Piske, Mackay, and Flege, 2001).

Nevertheless, many studies (e.g. Tahta et al., 1981a) show that (frequency of) L2 use is significantly associated with foreign accent: the more the L2 is used, the better is the pronunciation of the L2. For example, Flege, Munro and Mackay (1995) found that language use at work, at home, or with friends was the second major factor in accentedness (after AOL).[9]

---

[8] Adult speakers can also attain native-like pronunciation. In Ioup et al. (1994), two adult participants were rated as natives in the production and perception of Arabic. Obler (1989) also reports an exceptional speaker who learned several different languages after puberty and attained native-like proficiency. Finally, Bongaerts (1999) and Bongaerts et al. (1997, 2000) investigated L2 adult speakers of different L1 backgrounds, and reported that some speakers attained native-like pronunciation in sentence reading tasks and spontaneous conversation. It seems that there are some, but not many, such exceptional speakers. However, Birdsong (1999) claims that almost 30 % of his participants in French speech tests reached native-like proficiency, and we cannot ignore these participants as outliers.

[9] In earlier studies (Flege and Fletcher, 1992; Oyama, 1976), L2 use was not found to be a significant factor. Closer reading of Flege and Fletcher (1992), however, shows that reported L2 use is significantly correlated with judged accentedness of the speaker ($r = .431$) at the .05

Piske, Mackay and Flege (2001) showed that LOR was no longer significantly correlated with perceived accentedness when L2 use was partialed out in the analysis. The reason that LOR was not found to be a significant factor was speculatively accounted for as follows: First, after a certain age, the amount of input does not affect the L2 proficiency significantly, which supports the critical period hypothesis. Second, the amount of L2 use varies greatly among learners. Third, the quantity of input might not be as important as the quality of input. In Flege and Liu (2001), late Chinese bilinguals were cross-classified into a group with short LOR (less than 3.8 years) versus long LOR (more than 3.8 years), and a student versus non-student group. Participants took several tests, including an English stop identification test and a listening test. The results showed that long LOR only guaranteed success for students but not for non-students. Furthermore, the student group as a whole performed better than the non-student group.

The conclusion seems warranted, therefore, that experience with the L2 is indeed an important determinant of degree of accentedness in the L2. Length of residence and frequency of L2 use, however, are only rough statistical indicators of experience. More accurate predictions could probably be made if the details of the learning situation were taken into account.

### 2.2.3 Transfer from the native language

The pronunciation of sounds in adults' native language and the differences between those sounds often interfere with recognizing a foreign speech sound or distinguishing one foreign speech sound from another. Weinreich (1953) defines interference phenomena as "those instances of deviation from norm of either language which occur in the speech of bilinguals as a result of their familiarity with more than one language," and then adds that "the greater the difference between the two systems, the more numerous the mutually exclusive forms and patterns in each, the greater is the learning problem and the potential areas of interference." Since then interference has been attributed to the fact that between any two languages there are similarities and differences on all levels of analysis. As Weinreich implies that the degree of interference that would ensue from the partial similarities and the complete differences between the two competing categories, one is in the learner's native language and the other in the target language. Linguists assume that by comparing the relevant categories in L1 and L2 the area of interference between L1 and L2 can be predicted.

Linguists attribute the ease or difficulty of learning L2 phonological categories to (i) the competing phonemic categories of the L1 and L2 systems, (ii) the allophone membership of the phonemic categories and (iii) the distribution of the categories within their respective system (Brière, 1968).

Lado (1957) argues that there is a hierarchy of difficulties in learning the phonological categories of a foreign language. He defines the area of difficulties in terms of:

---

level, and even at the .01 level if one-tailed testing is accepted. So there seems to be no conflict between these and later publications on the topic.

(1) The distinctive versus the non-distinctive features of the two systems: Does the native language have a phonetically similar phoneme?
(2) The allophonic membership of the phonemes: Are the variants of the phonemes similar in both languages?
(3) The distribution of phonemes: Are the phonemes and their variants similarly distributed?

According to Lado's contrastive hypothesis, similar sounds are physically similar to those of the native language, that pattern similarly to them, and that are similarly distributed. These similar sounds will be easily learnt by simple transfer without any difficulty (positive transfer). On the other hand, sounds that are physically different from the L1 system, that structure differently, and that are distributed differently, will be the most difficult for L2 learners (negative transfer).

In the next three subsections we will summarize and briefly discuss three current views on transfer from the native (source) language to the foreign (target) language in so far as they relate to the acquisition of the L2 sound system. All three models address the issue to what extent foreign accent, and learning problems, can be predicted by comparing the sound systems of source and target language.

#### 2.2.3.1 Flege's Speech Learning Model (SLM)

By comparing the systematic similarities and differences between the actual pronunciations of foreign and native sounds, Flege (1987) defines L2 sounds which have no direct equivalent in L1 as "new" sounds and equivalent sounds which differ acoustically from their counterpart in L1 as "similar" sounds. Typically, a new sound is transcribed with an IPA basic symbol that differs from the symbol used to denote the equivalent sound in the native-language inventory. For instance, Dutch /ɛ/ is used as a substitute for the more open sound /æ/ in British English. After prolonged exposure to the foreign language, the learner will come to realize that the substitution is inadequate, and he will gradually form a new category for the foreign sound. Similar sounds are typically transcribed with the same base symbol from the IPA inventory and may differ only in diacritics (if at all). The auditory discrepancies between the native and foreign sounds are so small that the learner will never realize the substitution is harmful – even though his realizations of the target sounds may be noticeably incorrect when judged by native listeners of the target language. Flege's Speech Learning Model (SLM) predicts that the similar sounds are less easily produced and perceived in a native-like way than are new sounds, because the similar sounds in L1 have perceptual equivalence and merge into the same category in L2. This model predicts a greater (and more permanent) degree of difficulty for acquiring L2 sounds. The closer a target language sound is to the L1 sound, the more difficult it is to set up a new category for it (but also, the less the need for it, as the difference between source and target sound becomes negligible). Crucially, SLM makes the explicit claim that setting up new categories for the sounds in the L2 will go to the detriment of the categories in the L1, which will become less well-defined.

#### 2.2.3.2 Kuhl's Native Language Magnet model (NLM)

In her Native Language Magnet (NLM) model, Kuhl (1991) proposes that native language categories are prototypes. Each prototype occupies a specific location in a space whose dimensions are the phonetic properties that define that class of categories, as, for example, the vowel space is defined by vowels' formant frequencies. Tokens near a prototype are perceptually drawn towards it. This is why Kuhl refers to the prototypes as 'magnets'. Foreign as well as native sounds are drawn more strongly to these prototypes as a function of their proximity from them in the phonetic space. More distant foreign sounds either assimilate to another prototype if they are closer to it, or do not assimilate if there is no nearby prototype. Newly born infants come into the world with a fixed and large set of prototypes for all sort of vowels and consonants. As a result, sounds in the infant's language environment are perceptually attracted to some of the prototypes, and after six months or so certain prototypes have received ample reinforcement whilst others that have no function in the infants native language, have attrited due to lack of activation. In a sense, learning a first language is a matter of unlearning certain prototypes and at the same time tuning the activated prototypes. When at a later stage in life a second language has to be learned, the learner will assimilate the foreign sounds to the existing prototypes in his native language inventory, so that it is very difficult to perceive any difference between the foreign sounds and their equivalent native sounds, as they are all assimilated to the same prototype and therefore sound alike. The second-language learner's main task, then, is to set up new prototypes – by reactivating and tuning attrited prototypes between existing native-language prototypes to account for the foreign sounds. NLM holds that the new prototypes will be set up without degrading the prototypes that were already in place for the native language. Like SLM, NLM is primarily a perception-driven model of language learning.

#### 2.2.3.3 Best's Perceptual Assimilation Model (PAM)

Best analyses foreign accent in terms of similarity and difference between articulatory gestures across languages. Articulatory gestures refer to articulatory organs (active articulator, including laryngeal gesture), constriction locations (place of articulation), and constriction degree (manner of articulation). Different phonetic segments in different languages have different gestural constellations (in Best's terminology). Because all human languages draw upon the same set of gestural possibilities of the human vocal tract, there is usually a great deal of overlap among languages in the gestures and constellations contained within their individual phonological spaces, at least at segmental level. Non-native (foreign-accented) segments are those whose gestural elements or intergestural phasing do not match precisely with any native constellations (Best, 1995). Note that PAM does not appeal to perceptual characteristics of the sounds; perception is necessarily mediated through articulation, which makes the model a reincarnation of the motor theory of speech perception ('direct realism').

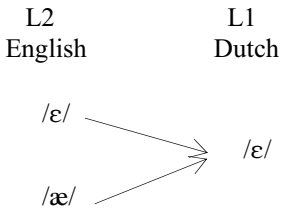
Focusing on these non-native segments, Best and co-workers try to determine to what extent the non-native segments are perceived in terms of the structures of the

source language, according to their similarities to, and discrepancies from, the native segmental constellations. For instance, if the listener's language has no ejective stops but does have voiceless aspirated and prevoiced stops, the glottal gestures and phasing of ejectives will be more similar gesturally to the voiceless aspirates than to the prevoiced stops, given that both the glottal closure and glottal widening prevent voicing and the glottal gestures of both are phased so as to reach their peaks at the release of the oral closure with which they are linked. Similarities between non-native segments and native gestural constellations determine the listeners' perceptual assimilation of the non-native phones to native categories. The listener is expected to detect gestural similarities and discrepancies to native phones. It is also expected that the listener will detect deviations from the gestural properties of native constellations as well. When the non-native sounds are very different from the native phones, they may be perceived only as having speech-like properties but may not assimilate strongly to any particular native category. In the extreme case they may not be recognized as speech (i.e. in terms of gestural constellations) but may instead be heard as non-speech sounds, such as clapping hands or flicking fingers. Best predicts that the non-native phones will be perceived in three ways:

(1) assimilated to a native category;
(2) perceived as an uncategorized speech sounds (this happens when the non-native phone falls in between two native categories (i.e. similar to at least native phones);
(3) perceived as a non-speech sound, which arguably happens when the non-native phone bears no resemblance to any phone in the native system.

PAM predicts several pairwise assimilation types. When non-native phones are phonetically similar to two different native phonemes and assimilate separately to them, the assimilation pattern is termed Two Category assimilation (TC). Flege calls it 'old contrast', e.g. the perception of English /$p^h$/ ~ /b̥/ and Spanish /o/ ~ /u/ by Dutch listeners (Escudero, 2001). In this pattern, the learner associates a binary contrast in the L2 with a binary contrast in L1. According to Best (1995) and Flege (1995), this leads to good category differentiation. Yet, this pattern may cause a perceptual problem, namely a boundary mismatch in the learner's L2 perception system, leading to problems with lexical access (Escudero, 2005). The learner should solve this problem by *shifting* the boundary between the categories in her L2 to match that of the target language.

Next, two sounds in the target language may, instead, assimilate equally well or poorly to a single native phoneme, termed Single Category assimilation (SC). Or both might assimilate to a single native phoneme, but one may fit better than the other, termed a Category Goodness difference (CG). An example of this assimilation pattern is given here:




Alternatively, one non-native phone may be Uncategorized as defined above, while the other is Categorized, forming an Uncategorized-Categorized pair (UC). Or both non-native phones might be Uncategorized speech segments (UU). Two phones' articulatory properties may both be quite discrepant from any native phonemes, and be perceived as Non-Assimilable (NA) non-speech sounds.

Discrimination of non-native contrasts can be hindered, aided or unaffected by native phonology, depending on how the non-native phones relate to native phonemes and contrasts. Native phonology should aid discrimination when the two phones are separated by native phonological boundaries, but should hinder it when both phones assimilate to the same native phoneme. TC and UC contrasts should be discriminated quite well, because in both cases the contrasting phones fall on opposite sides of a native phonological boundary. On the other hand, with the CG and SC types, both phones assimilate to the same native phoneme, so discriminability is impeded by native phonology. If one phone is good and the other is poor, discrimination will be very good (CG difference), but not as good as in a TC contrast because it is hindered by assimilation to a single native phoneme. In the SC case, both non-native phones are equivalent in phonetic goodness, hence discrimination is poor, hindered both by lack of phonological contrast and by lack of difference in fit.

PAM makes predictions about how listeners will recognize (or: assimilate) non-native phones with respect to the phonological categories of their native language, and how they will discriminate non-native contrasts.

When surveying the various models discussed above, it seems that it is very difficult to predict the specifics of foreign accent from a systematic comparison of phonetic and/or phonological properties of the source and target language. In spite of Flege's heuristic (same symbols different = new sounds, only diacritics different = similar sounds), learners sometimes have no problems where they are predicted or experience great learning problems where the model predicts their absence. Often, the researchers were honest enough to own up that their classification of sounds and sound contrasts in second language learning were based on existing pedagogical wisdom, and did not directly follow from any contrastive analysis. Given this unsatisfactory state of affairs, it might be worthwhile briefly considering an alternative option, which basically is a formalized procedure of being 'wise after the event'.

### 2.2.4 Alternative approach

Since a few years an alternative for contrastive analysis has been to use techniques from speech technology and apply these to the problem of foreign language learning. Automatic speech recognizers can be trained to recognize the sounds in the native language. Once the system is properly trained on a sufficient number of tokens of the target sounds (vowels and consonants) as spoken by a homogeneous group of first-language speakers, it will be able to successfully classify any new token as a sound in terms of the inventory of the target language, as long as the sounds are being spoken by a member of the same linguistic community as the training set. The recognizer technology as such is not directly relevant to the methodology; it may be Hidden Markov Modelling, Neural Networks, some hybrid mixture of both or even a more traditional multivariate statistical analysis such as Linear Discriminant Analysis. The point is that whatever the nature of the classification algorithm used, it can be applied to the task of classifying the sounds in another language that deviates to a greater or lesser degree from the training language. The algorithm will then misclassify the sounds in the foreign language by the same system of the training language, yielding crucial errors in the classification pattern. In fact, the confusion structure that we obtain in the vowels after they have been classified by the algorithm will be close to the misclassification we will obtain when foreign-language learners were asked to the same thing, i.e. the classification results/errors are a good prediction of the learning problems of human second language learners. This procedure has been used with reasonable success by Strange, Akahane-Yamada, Kubo, Trent, Nishi and Jenkins (1998) and by Strange, Bohn, Trent and Nishi (2004).

## 2.3 Measurement of intelligibility

### 2.3.1 Terminological preliminaries

Let us first consider what we mean by intelligibility. The exchange of ideas between speaker and listener has often been described in terms of a series of processes which together make up the so-called speech chain. When the stream of sounds impinges on the listener's eardrum, the listener will recognize linguistic units in the stream of sounds, viz. words, which appear in a particular order. A stretch is perfectly intelligible if all the words in the utterance are correctly recognized in the correct order. This is not the same as speech understanding (also: comprehension). Although the recognition of the words and their linear order is a precondition, comprehension is obtained if the listener correctly reconstructs the speaker's intentions, i.e. if the listener understands what the utterance means. A nonsensical utterance such as the beginning of Lewis Carroll's (1872) 'Jabberwocky':

> 'Twas brillig, and the slithy toves
> Did gyre and gimble in the wabe:
> All mimsy were the borogoves,
> And the mome raths outgrabe

is technically intelligible – as we can tell exactly what the words are (how many, and how they relate to each other) but we cannot reconstruct the writer's original intention as the words do not exist in the lexicon; it is therefore not comprehensible.

As a consequence of this we distinguish between the processing modules for word recognition (in sentential context) and higher-order integration of the word meaning into multi-sentence understanding (comprehension). Separate tests are required to test intelligibility and comprehension. Typically, intelligibility tests determine the number of correctly reproduced linguistic units (sounds, words) without ever asking if the listener understood the meaning of the utterance. Comprehension tests, on the other hand, ma check whether the listener has understood the meaning of utterances, for instance by asking the listener to choose whether the sentence is true, unlikely or nonsense (sentence verification). However, it is never an explicit concern for a comprehension test to check whether the listener has recognized all the words.

A second distinction we need to make is that between functional testing and opinion testing. Whether we are dealing with the testing of intelligibility or of comprehension, two types of test methodology are possible. Opinion tests ask the listener to subjectively rate a stretch of speech along one or many rating scales. For instance, an opinion test of intelligibility might ask the listener to assign a score to a foreign-accented utterance between 1 and 7 along a scale of intelligibility, where '1' would mean 'I think it is impossible to even recognize a single word' and '7' might mean 'I think it would be very easy to recognize all the words in this utterance perfectly'. Intermediate values would represent intermediate degrees of difficulty (i.e. intelligibility). Research has shown that native listeners have excellent and reliable intuitions on the relative intelligibility of (foreign-accented) speech utterances, with high within and between-rater agreement. Such a procedure may allow us to rank order foreign-accented utterances or speakers, but it will not tell us what the percentage of correctly recognized words will actually be. This is why we need functional tests. Functional tests require the listener to recognize words (when we are interested in intelligibility) or to actually grasp the meaning of the sentence(s) (when we are targeting comprehension). In our research we will only deal with functional tests of intelligibility. Nevertheless, in this introductory chapter we will briefly deal with both techniques for the testing of intelligibility as well as comprehension – but the emphasis will always be on functional testing.

When it comes to functional testing, a further split in tests has to be made in terms of on-line versus off-line techniques. On-line tests require the subject to respond when they are still processing the auditory stimulus, i.e. the process is tapped while it is still in full swing. Off-line tests allow the subject time to reflect before issuing the response. In the intelligibility tests used in this dissertation we only used off-line measures. This was not a principled choice but merely one based on convenience. Especially because a large part of the testing had to be done in the field (in China and in the United States), where we had no easy access to laboratories with possibilities to run on-line tests efficiently (using multiple work-stations), we were content to use more traditional off-line tests. These can be administered to small groups of listeners in parallel, without the need of sophisticated equipment.

### 2.3.2 Functional tests of intelligibility

As explained above, intelligibility involves the correct recognition of linguistic units in their linear order. Units exist at various levels of the linguistic hierarchy. The lowest level, with the smallest units, is that of the individual speech sounds or phonemes, i.e. the vowels and consonants. These are in principle units that carry no meaning of their own. This is a characteristic they share with consonant clusters. The smallest linguistic unit with a meaning of its own is the morpheme. Since morphemes often coincide with short words, we will collapse the word and morpheme levels for the purpose of the present study. In the following subsections we will briefly present test methods that have been devised to determine the intelligibility of speech at the level of the phoneme (consonants, vowels, clusters) and at the word level.

#### 2.3.2.1 Intelligibility of consonants (onset, coda)

When discussing techniques that have been developed for the testing of the intelligibility of consonants, we will not deal with perceptual experiments that target single contrasts. For instance, there is a very substantial literature on the perception of the /r/ ~ /l/ contrast in English by Japanese learners. Rather, we will survey techniques that determine the intelligibility of all the consonants in the target-language system, so that the results may be used diagnostically to determine which consonants, and which contrasts between them, are a learning problem and which ones are easy. Sounds in language naturally occur in the context of a word. So it would seem reasonable to present listeners with words containing the various consonants that make up the phoneme inventory of the target language. When doing so, however, one runs the risk that the (foreign) listener will correctly determine the identity of a consonant without actually having heard the sound correctly. This may happen as a result of lexical redundancy. For instance, when the listener hears the rhyme portion of a monosyllabic word such as /…ɔp/, only some consonants can be considered, viz. /p, t, k, m, ʃ, tʃ, h/ as in *pop, top, cop, mop, shop, chop* and *hop*, respectively. All other singleton consonants would not qualify as they do not combine with the rhyme /…ɔp/ to make up existing words in the English lexicon. As a result, getting the identity of the onset consonant in /…ɔp/ right would be a mixture of using bottom-up information provided by the consonant signal and top down information supplied from the lexicon. Unfortunately, there is not a single word frame in the English language that would allow each of the consonants in the language to appear in the onset position (let alone in intervocalic position). Therefore three other ways are commonly used to overcome the confound with lexical redundancy in intelligibility testing. The first is to present rhymes that allow only a small set of consonants to be filled in and list a closed set of printed alternatives (typically four) for the listener to choose from (MRT or Modified Rhyme Test). We refer to Van Heuven and Van Bezooijen (1995) for a more elaborate survey of testing procedures. Although this procedure does not eliminate lexical redundancy as a source of extra information, at least its effect is kept

constant. The second possibility is to have the full set of consonants embedded in a _VC frame, and force the listener to choose from the full inventory whether the response would be sense or nonsense. This creates the risk that the listener may have a lexical bias such that response alternatives that are words will be favored over alternatives that do not yield an existing word. The third possibility is to embed the consonants in fixed V_V structures that always result in nonsense items, so that the risk of lexical bias does not arise. In our experiments we have chosen for the latter option, as it is a highly efficient solution that does not involve the risk of lexical bias.

The intelligibility of consonants may differ substantially depending on their position in the syllable. It has been found that, generally, consonants in onset position are more difficult to identify correctly than the same consonants in coda position. One reason for this asymmetry might be that, across languages, the number of different coda consonants is smaller than the number of onset consonants. Mandarin Chinese, for example, has a set of 21 consonants which may appear in the onset, of which only the nasals /n/ and /ŋ/ remain as possible coda consonants. In Dutch and English the distribution is less lopsided but still asymmetrical: Dutch has 17 onset consonants against 11 in the coda, and English has 23 versus 21 (lacking /w, j, h/ but including /ŋ/, which is not an onset consonant), respectively. In our intelligibility tests of consonants we concentrated on the onset position.

#### 2.3.2.2 Intelligibility of vowels

When it comes to testing the intelligibility of vowels in languages such as English and Dutch, there is generally no need to resort to the use of nonsense items. As it happens, there are fully productive consonant frames that allow the insertion of any vowel in the inventory of the language and still yield a meaningful, existing word or phrase. The most widely used context for vowels in English is the /hVd/ frame. This frame was first used in the classical study by Peterson and Barney (1952), and has been used over and over again in later studies. Because of its wide-spread acceptance, we decided to follow established practice here, and adopt the same methodology.[10] This obviates the need for rather cumbersome and time-consuming tests such as the Minimal Pairs Intelligibility Test, which does test minimal vowel pairs in sentence context (Van Santen, 1993).

[10] In the classical study of Dutch vowel formants, Pols, Plomp and Van der Kamp (1974) used the /hVt/ frame. Given that Dutch has final devoicing, this seems a reasonable substitute for the English /hVd/ frame, were it not that the /hVt/ leaves several accidental gaps, so that the stimulus set is a mixture of sense and nonsense words. There is in fact only one fully productive consonant frame for Dutch, which is /rVt/ - thanks to the existence of proper nouns such as *Ruud* /ryt/ (short for Rudolph) and /rət/ *Ruth*). The problem here, however, is that the /r/ has many allophones (differing among other things in place of articulation, i.e. apical versus uvular) and that it is difficult to segment from the vowel.

#### 2.3.2.3 Intelligibility of clusters

The importance of consonant clusters in English should not be underestimated. About 40% of one-syllable words in English begin and 60% end with consonant clusters (Spiegel, Altom, Macchi and Wallace, 1990). The Bellcore Test and the CLID Test have been developed to fill this gap in the test batteries. The CLID Test (CLuster IDentification Test, Jekosch, 1994) is a very flexible architecture which can be used for generating a wide variety of monosyllabic stimuli (e.g. CCV, VCCC, CCCVVC) in an in principle unlimited number of languages, as long as matrices are available with the phonotactic constraints to be taken into account. Both the intelligibility of initial and final consonants and of (sequences of) medial vowels can be tested. In contrast to the CLID Test, the Bellcore Test (Spiegel et al., 1990) has a fixed set of stimuli, comprising both meaningless and meaningful words. Sequences of consonants are tested separately in initial and final position. The test has been applied to assess the intelligibility of two speech synthesizers compared with human speech presented over the telephone. The syllable score for human telephone speech was 88% correct.

In our experiments we only tested the intelligibility of consonant clusters in onset position preceded and followed by the vowel /ɑ/. We included 17 double consonant clusters /aCCa/ and supplemented these with three triple consonant clusters /aCCCa/. In this way the same format could be used as the one we employed in the case of single consonants. All the stimuli used were nonsense items. As a result, we did not test the intelligibility of clusters in coda positions, nor could we test for possible interactions of consonant articulation and the coarticulated vowel segments. It was felt, however, that the materials selected covered a sufficiently wide range of potential pronunciation problems of foreign learners of English. Including an even larger set of materials would have rendered the experiment unmanageable.

#### 2.3.2.4 Word recognition tests (on-line, off-line)

The (monomorphemic) word is the smallest unit in the language that links a sound shape with a meaning. We assume that words are stored in the mental lexicon, where they are specified, among other things, in terms of their sounds and the order of the segments them, rhythmic structure (number of syllables and the position of the stressed syllable, in so far as the language has stress), syntactic properties, and meaning. Intelligibility was defined above as the extent to which the words in an utterance can be recognized in the same order as they were produced by the speaker. Word recognition tests therefore play a prominent role in intelligibility testing.

Word recognition can be restricted to isolated target words that are presented without any spoken context. This is convenient for diagnostic purposes. If the listener fails to recognize the word, the problem should be located in the word itself. However, words in everyday communication hardly ever occur in isolation. Therefore, word recognition in connected speech is a more realistic test. When the recognition of a target word fails, however, it is difficult to determine whether the cause is in the target word itself or whether the failure is due to the fact that some earlier words were not recognized and failed to constrain the identity of the later

target word. A solution to this dilemma has been found in presenting the same target both in context and in isolation. This is a laborious solution, however, as the stimuli must be blocked over two groups of listeners (who have to be equally proficient) in order to prevent learning effects (priming).

In our experiments we used two types of word-recognition tests. The first used so-called Semantically Unpredictable Sentences (SUS-test, Benoît, Grice and Hazan, 1996), in which simple (often monosyllabic) words were presented in sentences but where the sequence of words made no sense, as in *The state sang by the long week* (for more details on the SUS test, see § 4.2.4). The content words are not made more predictable by earlier content words in the utterance, so that the test is actually an accumulation of single word recognition items. The second word recognition is the Speech In Noise (SPIN) test. This test requires listeners to recognize sentence-final words which are or are not predictable from the earlier context (Kalikov, Stevens and Elliot, 1977). This test will be discussed below in § 4.2.5. Both tests require the listener to write down the target words by way of dictation. There are no severe time constraints on the task performance, so that these are basically off-line word recognition tests, which inform us to what extent the word recognition was a problem for the listener, but disclose nothing about the ongoing word recognition process.

There are several on-line techniques that tap the word recognition process in real time. Since we chose not to use on-line techniques (see above) we will be brief about them. Most on-line techniques require listeners to detect the presence of some feature of a word, by pressing a response key as quickly as they can manage. The response time is indicative of the moment the word was recognized. The features to be detected can be of several kinds. In phoneme detection, listeners are instructed to press a button as soon as they detect its presence in the stimulus. Generally, each individual sound making up a word is acoustically unreliable. Therefore listeners wait until they have recognized a word in a sequence of sounds. Phoneme detection is therefore indicative of the time it takes to recognize the word that harbors the target. The more difficult the word is to recognize, the longer the target detection will take. Alternatively, listeners may be instructed to detect the presence of some semantic feature, e.g. press a button when they hear the name of an animal, or when they hear a word that expresses a tangible object (rather than an abstraction). Here the rationale behind the test is that the listener may only access the meaning of a word in the mental lexicon after the word has been recognized, so that semantic property detection again is indicative of word recognition time. An interesting alternative that was found to discriminate quite clearly between native and non-native listeners (Poelmans, 2003) is the lexical decision technique. Here the listeners are presented with sound sequences that either constitute a word or a nonword. The subject is instructed to press one of two buttons marked ‘word’ or ‘nonword’, depending on his decision. Obviously, the decision that the stimulus is a word can only be made once the word is recognized, so that this task, too, indicates the time needed (and thereby the difficulty) for word recognition.[11]

[11] Interestingly, discrimination of native and non-native listeners was most clearly achieved in the (correct) rejection of non-words rather than in the (correct) acceptance of words in Poelmans (2003).

### 2.3.3 Functional tests of speech understanding (comprehension tests)

Listeners have understood (or: comprehended) a sentence or longer stretch of speech if they have grasped the meaning of the passage. There are several functional tests that have been employed to test the quality of the listener's understanding or comprehension. Broadly, these methods can be of four types: (i) having listeners answer questions on the contents of the passage, (ii) verifying the truth of sentences, (iii) verifying on-line descriptions of still pictures or video footage, and (iv) carrying out spoken instructions.

*Comprehension questions*. Either before or after the stimulus speech passage is presented, listeners are given a specific question that can be correctly answered only if they understood the contents of the passage. Asking the question before presenting the passage diminishes memory load and tests intentional listening. Asking the question post hoc makes heavier demands on memory and may therefore be less desirable. The comprehension questions can be of the open or closed type. Open questions ask listeners to formulate and write down their answers from scratch, closed questions present the listeners with two, three or four alternative answers from which they have to choose the correct one. Answering comprehension questions are off-line tests. Listeners have ample time to think or recall the speech and give the answer.

*Sentence verification tests*. Here listeners are asked to judge whether a sentence they hear, expresses a truth, is unlikely, or nonsense, by pressing one of two keys marked 'true' or 'false' immediately after they hear a sentence. For instance, a stimulus *People wear their hats on their feet* should be responded to by pressing 'false'. In this kind of test, listeners can make the right choice only if they understood the sentence correctly.[12] Sentence verification can be used as an on-line comprehension testing technique. In order to do so, subjects should be asked to press the response key as fast as they can manage, if possible even before the end of the speech utterance has been reached.

*Descriptive language*. One way to test comprehension of speech is to ask a listener to indicate whether a spoken description does or does not match a visually presented scene, by pressing one of two keys marked 'correct' or 'wrong'. The visual presentation can be a in the form a still picture or it can be a scene from a movie. By aligning the spoken description with the development of the scene (only possible in video footage) on-line comprehension can be tapped.

*Carrying out instructions*. A last test we want to mention here relies on carrying out spoken instructions. The listener is asked to carry out certain actions following spoken instructions recorded on tape. Obviously, listeners can only carry out the instructions if they understood them. The implementations of this technique range from crude to sophisticated. A crude but effective use of the technique is the Token Test, which has been in service for decades to test speech understanding with patients suffering from brain lesions. The patient has several geometric objects on

[12] Given just two response alternatives (true, nonsense) the chance of getting the correct response by guessing is 50%. The role of guessing quickly diminishes with the number of items in the test. When a binary verification test comprises 50 items, the chance of answering all items correctly by pure guessing is very small.

the table in front of him (circle, square, triangle) which may have different colors. Instructions are of the type: 'Put the red square on top of the yellow triangle'. Van Heuven (1986) describes a similar technique used to determine differences in comprehensibility of several types of deviant (foreign-accented) Dutch speech, and shows that the technique is very sensitive. More recent versions of the technique no longer require the physical manipulation of objects in space but instruct the listeners to manipulating objects on a screen by moving a joystick or computer mouse, which also affords easy measurement of response time. The instruction tests can be conceived of as on-line tests.

Comprehensibility and intelligibility have been found to be related with global foreign accent scores, but since we will test intelligibility rather than comprehension, we will not further discuss comprehension tests in the present study.

### 2.3.4 Information reduction techniques

Especially when the intelligibility of speech produced by native speakers is heard by native listeners, performance levels tend to be close to ceiling, so that small differences in proficiency between speakers or between listeners are hard to detect reliably. An often used solution to ceiling effects is the use of information reduction in the stimulus. The underlying idea is that speech is a highly redundant code, and that native listeners may use the redundancies better than non-native listeners. There are several signal degradation techniques that can be used to reduce the redundancy in the spoken word forms. First, we may obscure the speech signal with noise so that some of the distinguishing properties of the word are no longer audible. Second, we may eliminate certain frequencies or frequency bands from the signal, such that important distinguishing frequencies are no longer available. Third, we may simply eliminate complete segments or larger parts of the speech signal by replacing them by stretches of silence or by noise, without changing the temporal relationships among the sounds that remain. We will briefly discuss these techniques and review how they have been used in the study of intelligibility of (foreign-accented) speech.

#### 2.3.4.1 Speech in noise

One of the earliest applications of speech in noise has been the development of testing materials for audiological purposes. In order to determine the extent of hearing loss with hard-of-hearing patients the threshold of hearing may be determined by asking listeners to recognize words presented to them in noise. On the first pass the noise is much stronger than the speech, so that the word cannot be recognized, not even by a healthy listener. On successive following passes the noise level is reduced in steps of, say, 3 decibels, until the spoken word is loud enough – relative to the reduced noise level – to be recognized. The signal-to-noise ratio (expressed in dB) at which the word can be recognized is the intelligibility threshold. Using this measure, differences in intelligibility of different words, spoken by different speakers, or heard by different individuals (whether native or non-native, whether hearing-impaired or healthy) can be determined. A well-known

set of audiology sentences to be presented in noise was developed for the SPIN test (Speech In Noise) by Kalikow et al. (1977). In this test, simple monosyllabic target words were presented at the end of simple short sentences, which came in two varieties. In one type of sentence the final target word was used in citation form, such that its identity was not constrained by the preceding context, as in *They are discussing the (map)*. In the other condition the target word was strongly constrained by the context as in *She cooked him a hearty (meal)*. Results show that the predictable words were recognized at more severe signal-to-noise ratios than the unpredictable words. The same techniques, and even the same test materials, have later been used to test differences in intelligibility of synthetic speech (Van Bezooijen and Van Heuven, 1997 and references therein) and non-native speech production and perception (e.g. Van Wijngaarden, 2001).

It has been shown that speech is more or less effectively masked depending on the specific type of noise. For instance, white noise (in which all frequencies occur with equal chance and with equal amplitude) is a less effective masker than noise which has roughly the same spectral distribution as speech, i.e. with emphasis on the lower part of the spectrum. Thus, pink noise (−3dB/octave) and even more strongly, ANSI noise are more effective maskers. The most effective type of all is so-called speech noise (also called babble noise) which is actually speech produced by the same speaker as the individual who spoke the target stimulus, dubbed several times with different phasing, see Eggen (1989). A second parameter in using speech in noise is whether the noise has constant intensity (so that loud sounds exceed the noise but low-intensity sounds – typically consonants – are completely masked) or whether the noise is modulated so as to follow the intensity contour of the speech stimulus.

In our experiments we used SPIN sentences but presented them without any added noise. As it turned out, the quality of the non-native speech (and of the non-native listeners) was so poor that the intelligibility was evenly distributed in the 30 to 90-% range.

#### 2.3.4.2 Filtering

It is a well-known phenomenon that a foreign speaker may successfully communicate with a native listener under ideal circumstances but that communication tends to break down when the telephone is used. The reason for the breakdown is that the telephone filters speech such that only frequencies in the restricted band between 300 and 3300 Hz are transmitted. The impoverished signal contains enough information for successful communication between two native speakers, who know the code, but when either the speaker or the listener is foreign, the signal is too poor to allow full intelligibility. For this reason filtered speech (high-pass, low-pass and band-pass) has been used as a means of degrading the speech stimulus in an attempt to make fine-grained determinations of differences in intelligibility of various types of materials; for instance French and Steinberg (1947), Hirsh, Reynolds and Joseph (1954), Miller and Nicely (1955), and others after these pioneering studies, used filtering in intelligibility testing.

### 2.3.5 Gating

The classical gating study was done by Grosjean (1980). In this study listeners were first presented with a short initial portion of a target word and asked to guess what the word would be. On second and later passes the audible portion of the target word was lengthened by one phoneme at the time, until the listener could supply the entire word. Highly frequent words and words in more constraining preceding contexts could be finished from short onset fragments than low-frequency words in less constraining contexts. Nooteboom and Truin (1980) used the same technique and showed that native Dutch listeners could recognize target words from shorter onset fragments than non-native (English) listeners of Dutch. Smeele (1985) showed that native-accented Dutch words were recognized by Dutch listeners from shorter onset portions that German-accented Dutch words.

Variations on the gating paradigm abound. Instead of suppressing the final portion of the target word, some researchers have suppressed the initial portion (replacing it by either silence or by noise – the latter condition proves more conducive to intelligibility). Moreover, the portions of the signal that are suppressed (zeroed) or replaced by noise need not be contiguous.

## 2.4 Can higher-order performance skills be predicted from lower-order components?

Sentences are made up of words, and words are made up of syllables, which in turn are composed of vowels and consonants. From the perspective of the speaker, being able to pronounce the sounds in a word correctly is a skill that has relatively little to do with the skill of arranging the words to form a syntactically appropriate sentence. Pronunciation would seem to involve a good deal of motor skills, whilst arranging the morphology and syntax is a much more cognitive skill. In previous sections we discussed the notion of a critical period which allows the formation of native or near-native pronunciation skills in a second language. No mention is ever made of a critical period needed for the acquisition of the morpho-syntax of a second language. This in it self would seem to suggest that the lower-level phonetic skills have little in common with the higher-order syntactic skills. On the strength of this argument we would expect a weak correlation between lower and higher-order skills when it comes to speech production in a second language.

When we approach the problem from the perspective of the listener, the argument will be different. It would make sense to predict that words with poorly articulated sounds will be hard to recognize, and sentences made up of poorly recognized words will not be understood. Accordingly, we may ask how well word recognition scores can be predicted from the consonant and vowel identification scores obtained for the same speaker. If the sounds can be successfully identified then we expect the word recognition scores for the same speaker to be high, too. We may also ask the question which of the two sets of sounds would be the better predictor of word recognition, vowel identification or consonant identification scores.

In order to answer such questions one needs data for a fair number of speakers, who range between poor and good. In the data to be collected in the present project we recorded speech from groups of 20 American, 20 Dutch and 20 Chinese speakers of English, who pronounced vowels, consonants and words in context. Vowel and consonant identification scores were obtained for all 60 speakers but only with a subset of the materials. Full vowel and consonant identification scores as well as word recognition scores were collected from only six speakers (two Chinese, two Dutch and two Americans). No word recognition was tested for the remaining 54 speakers. Given the extremely small number of speakers, trying to predict word recognition in SUS sentences and in SPIN sentences is hazardous. We will nevertheless present an analysis of the relationships.

## 2.5 Problems at the phonological level vs. phonetic level

When analyzing problems in the acquisition of the sound system of a second language, it has been customary to distinguish phonological problems from phonetic ones. It is not always clear if there is a boundary between these two disciplines, and if the distinction is useful at all when applied to foreign language acquisition. We will make an effort to separate the two, and consider what the distinction may contribute to our understanding of the problems.

### 2.5.1 Phonology

By phonology we mean the abstract structure of the sound system of a language, in the abstraction of the precise phonetic implementation of the sound categories. Properties that can be studied in terms of abstract structure are the number of sounds in the inventories of source and target languages, the distinctive features needed to organize the sounds in the inventory in contrasts (oppositions) and the constraints on the formation of legal syllables. Differences between positional allophones within the same phoneme category are also subsumed under the heading of phonological structure.

#### 2.5.1.1 Differences in inventories (number of sounds, oppositions)

Obviously, two languages differ in their sound system if there is a difference in the number of sounds between the two languages. In our research we will study the type of English spoken by speakers with Dutch and Chinese as their native language. The number of phonemes in English, Dutch and Chinese differ considerably. Generally, the Dutch inventory is more like the English system than the Chinese system is. For instance, in both Dutch and English the number of vowels is much larger than in Chinese. In order to divide the vowel space among the vowels in the inventory, English and Dutch vowels must differ along more parameters than the vowels in the Chinese inventory. Typically, then, Dutch and English vowels are differentiated in addition to other parameters, by their length (or tense~lax). Chinese does not employ a length (or tense~lax) contrast, so that presumably Chinese learners of English will have at least one problem to overcome that should be easy for Dutch learners, i.e. learning to use the length contrast in English. In Chapter three we will present an

overview of the inventories of the three languages involved in the present study, and indicate by what articulatory features the various contrasts in the systems can be accounted for. We will make an effort to predict what kind of learning problems will be seen in the results of our experiments, following the models of Flege (SLM) and Best (PAM).

#### 2.5.1.2 Differences in syllable structure (no clusters, simpler clusters, no coda)

Even if we know the size and internal structure of the phoneme inventories of source and target languages, there are quite a few systemic properties that we cannot yet deal with. Languages differ widely in the way they build syllables from vowels and consonants. The simplest syllable type across languages is composed of a consonant (C) followed by a vowel (V). More complex syllable types can be formed either by omitting the initial consonant or by augmenting the number of consonants in the onset (i.e. preceding the vowel) or in the coda (i.e. following the vowel). Chinese has a predilection for simple syllables, whereas English and Dutch allow rather complex syllable types with clusters of multiple consonants in onset and/or coda.

Syllable types across languages are implicationally ordered, such that all the simple syllable types allowed in strongly constrained languages (such as Chinese) will also occur in less constrained languages with many complex syllable types (such as English and Dutch) but not vice versa (Blevins, 1985). Therefore, learners with a constrained source language will have a problem in producing the more complex syllable types in a less constrained target language. So we predict that Chinese learners of English will experience considerable problems when having to produce (and perceive) the complex English syllable types. Typically clusters will be broken up into several syllables by inserting epenthetic vowels between consonants that are not legal clusters in the source language. Dutch is probably a less constrained language than English, allowing more and more varied consonant clusters both in the onset and in the coda, so that generally we do not predict any problems with English syllable types for Dutch learners.

#### 2.5.1.3 Positional allophones (final devoicing)

In the preceding subsection we introduced the difference between onset and coda. It happens very often that a language uses clearly distinct allophones for the same phoneme such that one allophone occurs only on the onset position whereas the other appears in the coda only. English /l/ has two varieties, called clear (or light) and dark (also: dull or velarized). The clear /l/ is restricted to onset positions whilst the dark allophone is bound to coda positions. In this respect English and Dutch behave identically, so that we predict no learning problem here for Dutch learners of English. Chinese learners of English have a double problem. First they do not normally allow syllable types with a coda consonant – the only exception being codas with a nasal in them – so that learning to pronounce an /l/ in the coda is not only a problem as such, but also one that is aggravated by the fact that Chinese learners will have to learn to pronounce a clear /l/ in the onset and differentiate this from a dark /l/ in the coda.

A second, related problem is that certain sounds may occur on the onset that are not allowed in the coda. As a case in point, Dutch has an opposition between voiced and voiceless obstruents, as does English, but only in onset position. The

voiced~voiceless opposition is neutralised in coda position, where only the voiceless member may occur. In English, however, voiced obstruents abound in coda position. This presents a great learning problem for Dutch learners of English. Even though the source language has both voiced and voiceless obstruents in the onset position, Dutch learners find it very difficult to pronounce a voiced obstruent in the coda, as they should when they speak English. Apparently, then, sound contrasts do not generalise to other syllable positions than those they occupy in the source language. The Chinese learning situation presents an interesting test case. Chinese has voiced and voiceless obstruents in the onset, just like Dutch and English). However, Chinese has no coda consonants at all (with the exception of the nasal coda) so that, possibly, the Chinese learner of English is not impeded in setting up the voiced~voiceless contrast in the coda. It is unclear if this structural property would give the Chinese learner an advantage over his Dutch counterpart when learning English.

### 2.5.2 Phonetics

What is left for phonetics is the implementation of the categories and the contrasts between them. If two systems, say in source and target language, have the same number of sounds, organized in terms of the same oppositions, there may still be considerable differences in the division of the articulatory space between the various categories, and in the way the boundaries between the categories are cued by acoustic features

.

#### 2.5.2.1 Same oppositions but different boundaries

Source and target languages may have the same opposition along the same parameter with the same number of categories along the parameter, and yet differ in the phonetic implementation of the contrast. As an example, consider the phonetic implementation of the tense~lax contrast in stops in English, Dutch and Chinese. In each of the three languages stop consonants in the onset fall in one of two phonological categories, probably best characterized as tense (fortis, voiceless) as opposed to lax (lenis, voiced). The phonetic parameter that carries the contrast is often called Voice Onset Time (VOT). This is a complex parameter in that it is defined as the time interval between the moment of the release of the stop consonant (coinciding with a short noise burst and a sudden increase in energy at higher frequencies) and the moment the vocal cords start vibrating. In Dutch the lax member of the contrast has negative VOT, meaning that the vocal cords begin vibrating some 50 ms before the stop is released, i.e. there is glottal pulsation while the mouth is closed (‘prevoicing’). The Dutch fortis member has 0 VOT, so that glottal pulses are produced as soon as the mouth opens. The English tense~lax opposition is phonetically different. Here the lax member has 0 VOT, while the tense counterpart has positive VOT, meaning that the vocal cords do not start vibrating until well after the consonant is released. The time interval between the release burst and the onset of glottal pulsation is filled up with a whispered (voiceless) vowel sound called aspiration. From this difference in phonetic implementation we predict that Dutch speakers of English have a problem: they will use the Dutch implementation of the contrast when speaking English, with the result

that an English listener may well mistake the Dutch /p, t, k/ for English /b, d, g/, respectively, since all these sounds have 0 ms VOT. Chinese has roughly the same phonetic implementation of the tense~lax opposition in onset stops as in English, so that Chinese speakers of English would not have any problems here.

#### 2.5.2.2 Different cue tradings

Typically a phonological contrast is phonetically cued not along a single acoustic parameter but along multiple parameters simultaneously. For instance, the difference between tense and lax members of vowel contrasts in Dutch and English are cued by differences in duration (the tense members are some 50% longer than their lax counterparts) as well as by differences in vowel quality (timbre). Typically, the lax vowels assume more centralized positions in the articulatory vowel space than the tense counterparts. In German, however, the difference between the members is cued only by a difference in duration while the vowel quality differences are negligible (Strange et al. 2004). To compensate for the lack of a qualitative difference, the German tense and lax vowels have a larger difference in duration than in English. English speakers of German may apply the English implementation of the tense~lax contrast so that the durational difference in their German vowels is too small to be effective, while the Germans will be relative insensitive to the quality differences in English-accented German. A similar problem was noted by Van Heuven (1986) in the perception of Dutch vowels by Turkish learners. The Turkish learners revealed a very clear tense~lax contrast in the Dutch open vowel pair /a: ~ ɑ/ but were sensitive only to the duration cue and failed to pick up the quality cue (the tense member have higher first and second formant values than the lax counterpart) and consequently misclassified about half of the vowel tokens.

## 2.6 Concluding remarks

The purpose of this study is not to carry out any systematic testing of theories of second-language acquisition of sound systems. The above survey of such theories was merely offered as background information enabling the reader to understand why certain choices were made in the experimental part of my project. Nevertheless, whenever my results give rise to reflection on theoretical positions, I will do so and refer back to sections of the present chapter.