SPEECH INTELLIGIBILITY OF NATIVE AND NON-NATIVE SPEECH

Sander J. van Wijngaarden
E-Mail: vanWijngaarden@tm.tno.nl
TNO Human Factors Research Institute.
P.O. Box 23,
3769 ZG Soesterberg,
The Netherlands.

ABSTRACT

The intelligibility of speech is known to be lower if the talker is non-native instead of native for the given language. This study is aimed at quantifying the overall degradation due to acoustic-phonetic limitations of non-native talkers of Dutch, specifically of Dutch-speaking Americans who have lived in the Netherlands 1-3 years. Experiments were performed using phoneme intelligibility and sentence intelligibility tests, using additive noise as a means of degrading the intelligibility of speech utterances for test purposes. The overall difference in sentence intelligibility between native Dutch talkers and American talkers of Dutch, using native Dutch listeners, was found to correspond to a difference in speech-to-noise ratio of approximately 3 dB. The main contribution to the degradation of speech intelligibility by introducing non-native talkers and/or listeners, is by confusion of vowels, especially those that do not occur in American English.

1. INTRODUCTION

Many attributes of individual talkers are known to influence human speech intelligibility. Some of these are at the linguistic level (such as syntactical and lexical aspects [1,2]), some are at the acoustic-phonetic level (e.g. syllabic rhythm and speed, F_0-range, intonation, articulation of different phonemes [3,4,5]). Non-nativeness of a particular talker or listener may be interpreted as a specific category of attributes influencing speech intelligibility.
Among the attributes known to be related to non-nativeness of talkers are vowel-onset time, intonation, speaking rate and phonemic repertoire [e.g. 6,7]. Many fine-grained phonetic studies of second-language talkers have given insight in factors that may contribute to recognition of foreign accents [e.g. 8]. Also, factors contributing to speech intelligibility by non-native *listeners* were investigated [9,10]. Development of accents with experience in using a foreign language has been studied extensively [eg. 11]. Relatively much work has been done in the field of second language (L2) speech perception; however, many studies have been focussed on particular phonetic attributes or phenomena, usually across two (or few) languages.

An important motivation to study the effect of non-native speech, is the effectiveness of human speech communication. From this perspective, it is not important to have detailed knowledge of speech production by L2 talkers; it is more interesting to quantify the effect on the overall speech intelligibility in general terms.
This may be achieved by carrying out speech intelligibility experiments with L1 and L2 subjects (talkers/listeners) in a certain language, in our case Dutch. As with all speech intelligibility tests, a choice has to be made of test fragments: sentences, words or phonemes. In the case of words, meaningful words or nonsense-words may be used. Also, the paradigm will have to be suitable for non-native subjects; on one hand, the limited control of a second language is the object of study, on the other hand it may be experienced as a problem in carrying out some types of speech intelligibility tests (for instance those depending on typing out nonsense words by second-language listeners, who will have a tendency to use native-language spelling of some nonsense words).

2. EXPERIMENTAL SETUP

2.1. Test types

Two types of speech intelligibility experiments were performed: a sentence intelligibility test and a phoneme-intelligibility test based on nonsense-words. The sentence intelligibility test was essentially identical to a standard and widely used test method known as the Speech Reception Threshold (SRT) method [12]. The phoneme intelligibility test is closely related to the equally-balanced CVC test [13].

2.2. Speech Reception Threshold (SRT) method.

The sentence intelligibility test was a standard Speech Reception Threshold (SRT) experiment [12]. This test gives a robust measure for sentence intelligibility in noise, corresponding to the speech-to-noise ratio that gives 50% correct response of short redundant sentences.
In the SRT testing procedure, masking noise is added to test sentences in order to obtain the required speech-to-noise ratio. The masking noise spectrum is equal to the long-term spectrum of the test sentences. After

presentation of each sentence, a subject responds with the sentence as he or she perceives it, and the experimenter compares the response with the actual sentence. If the response is completely correct, the noise level for the next sentence is increased by 2 dB; after an incorrect response, the noise level is decreased by 2 dB. The first sentence is repeated until it is responded correctly, using 4 dB steps. This is done to quickly converge to the 50% intelligibility threshold. By taking the average speech-to-noise ratio at the ear over the last 10 sentences, the 50% sentence intelligibility threshold (SRT) is obtained.

During the actual experiments, the subjects (listeners) were seated in a sufficiently silent room. A set of Sony MDR-CD770 headphones were used to present the recorded sentences, diotically, to the listeners. Using an artificial head, distortion components introduced by the experimental setup were found to be sufficiently small.

2.3 Semi-open response equally balanced CVC test method

A type of semi-open response CVC (consonant-vowel-consonant) intelligibility test was developed for the purpose of testing phoneme intelligibility with non-native subjects. Using this test, recognition of initial consonants and vowels could be scored, and confusion matrices could be composed [14]. The method is similar to an open-response equally-balanced CVC paradigm [13]. The main differences are that the final consonant is not tested, and that the subject responds by choosing an alternative from a (nearly) exhaustive list of possible CVC-words, instead of typing the word in response to the stimulus. The advantage of this approach is that extensive training of subjects becomes unnecessary, while the construction of confusion matrices is still possible. Problems that were expected using a 'difficult' open-response paradigm with non-native subjects were successfully avoided.

During each 3 to 4 minute test, all test phonemes were tested once. Initial consonants and vowels with a frequency of occurrence (based on a Dutch newspaper) below 2% were not included in the test, leaving 17 initial consonants and 15 vowels. Thus, when testing an initial consonant, 17 alternatives were displayed on screen, and for a vowel 15 alternatives. When testing the vowel /Ø…/, for instance, the list of CVC words for the listener to choose from could be 'jaap', 'jup', 'jeup', 'jip', etc.

In each test, the order of presentation was randomized. The other phonemes in the CVC words, not tested themselves, were selected. Four of these non-tested phonemes, influencing the test through co-articulation effects, were selected per test, in an attempt to maximize the spread of these phonemes over a perceptual space [15]. Several selections of four non-tested phonemes were used for each talker.

2.4. Collection of speech material

The speech material was collected using a B&K type 4192 microphone with a B&K type 2669 microphone pre-amplifier. The sound was digitized using the wave-audio device of a Topline 9000 notebook-computer, which was screened for adequate bandwidth, dynamic range and electronic noise properties This same notebook-computer (with the same audio-device) was used to implement the test procedure.

Since non-native talkers of the Dutch language, matching all criteria, are rather difficult to find, the arguable choice was made to record the material at a location of the talker's choice. This proved to be an effective measure to facilitate the recruitment of subjects, but lead to a lesser control of the influence of background noise and room acoustics in the recorded material. To limit this influence, the microphone was placed at relatively close range (15 cm). Signal-to-noise ratios were verified to be always higher than 20 dB for all frequencies relevant for speech perception. Hence, no effects of the variation in acoustics and background noise on the outcome of the perceptual experiments is expected.

All speech material was calibrated to have the same speech level for each utterance. In the case of the CVC test, the utterance over which the speech level was determined was not just the CVC-word itself, but also the carrier sentence in which it was embedded.

2.5. Subjects.

Two groups of talkers were recruited, each group consisting of four subjects, two male and two female. The L1 group of talkers consisted of native talkers of the Dutch language without strong regional accents. The L2 group of talkers were native Americans, speaking Dutch fluently but with an accent that was immediately recognized by most listeners.

Perception and production of foreign speech sounds depends on the experience of subjects with the foreign language [11]. Also, the age of acquisition is of importance, leading to a distinction between early and late bilinguals. Generally, the transition age between those categories is found roughly to be puberty [eg. 11,16]. Three of the four L2 talkers had acquired knowledge of the Dutch language above age 23, and spoke Dutch for less than 3 years. The fourth subject (referred to later on as subject L2F8) had first learnt Dutch at age 13 and had been speaking Dutch for 18 years. Although this fourth subject, the only subject that might be categorized as 'early bilingual', showed appreciably better control of the Dutch language, the American accent was still readily noticed.

The L1 talkers were selected to match the L2 group in terms of age and level of education.

The L2 *listeners* all had over 12 years experience with the Dutch language (average 20 years), and used the Dutch language frequently in communication at home or work. No special requirements were included in the selection of the L1 listeners.

None of the subjects suffered from speech or hearing impairments, or any unusual hearing loss likely to affect the outcome of test results.

3. RESULTS

3.1. Sentence intelligibility

Four sets of sentence intelligibility experiments were carried out, corresponding to all combinations of L1 and L2 listeners and talkers. The condition with L1 listeners and L1 talkers may be seen as a baseline condition, involving only Dutch subjects. In figure 1, average results are given for these four conditions.

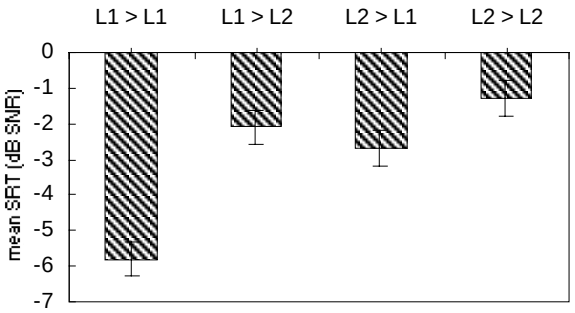

Figure 1. Results for four types of talker-listener combinations (16 talker-listener pairs per condition, mean values and standard errors given). L1 > L2, for instance, means native talker, non-native listener.

The lowest (most negative) SRT value is, as expected, for the baseline group with both L1 listeners and L1 talkers. This means that in this condition the highest noise level may be allowed to still obtain 50% correct sentence responses, down to a speech-to-noise ratio (SNR) of –6 dB.

The condition with L2 talkers and L1 listeners requires a 3 dB lower noise level for the same 50% sentence intelligibility than the L1>L1 condition. The L1>L2 condition (L1 talkers, L2 listeners) also allows less noise for 50% sentence intelligibility; the difference is now nearly 4 dB. The L2>L2 condition, showing the lowest intelligibility results, allows for 4.5 dB less noise.

Figure 1 gives us a general image of the influence of non-nativeness of speakers and listeners on speech intelligibility, at least for these particular L1 and L2 languages. It also shows that, even though the L2 talker group was less experienced than the L2 listener group, having L2 listeners gives relatively more degradation of speech intelligibility than having L2 talkers. The combination of L2 listeners *and* L2 talkers gives an additional degradation which is less than the degradation caused by L2 talkers and L2 listeners separately.

The results of figure 1 are also given in figures 2 and 3, but now by talker instead of talker/listener group. For the L1 listener group (figure 2), all L1 talkers offer better intelligibility than any L2 talker, although the difference between talker L1F4 and L2F8 is not significant. Figure 3 is quite different; to L2 listeners, the highest intelligibility is offered by one of the L2 talkers. The average score by L2 talkers as shown in figure 1 is quite low, but mainly because of talkers L2M5 and L2F6. The difference between L1 and L2 listeners is not as clear with L2 talkers as with L1 talkers.

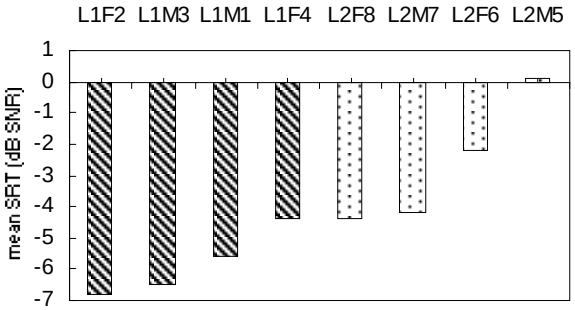

Figure 2. Mean SRT scores for eight individual talkers, with the L1 group of listeners (4 listeners per condition). L2M5, for instance, means L2 talker, male, talker #5.

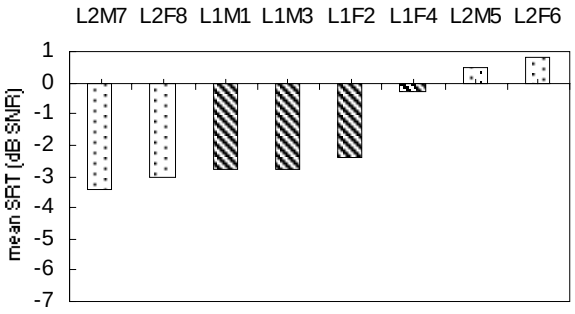

Figure 3. Mean SRT scores for eight individual talkers, with the L2 group of listeners (4 listeners per condition).

3.2. Phoneme intelligibility

The CVC-based phoneme test, although somewhat different in nature, may be expected to yield results that correspond well with the SRT results. However, the CVC test scores are percentages of correctly recognized phonemes, whereas the SRT results are speech-to-noise ratios to obtain 50% sentence intelligibility. To verify correspondence between both test types, CVC experiments were performed at various signal-to-noise ratios. Results, for initial consonants and vowels separately, are given in figures 4 and 5.

Due to the relatively small number of listeners, the experiment data are slightly too noisy for a clear polynomial curve fit. The general trend, however, may well be observed from the data.

At relatively low speech-to-noise ratios, the L2 talker leads to better initial consonant recognition than the L1 talker. At higher speech-to-noise ratios the initial consonant recognition of the L2 talker appears to saturate at a somewhat lower level then the initial consonant recognition of the L1 talker.

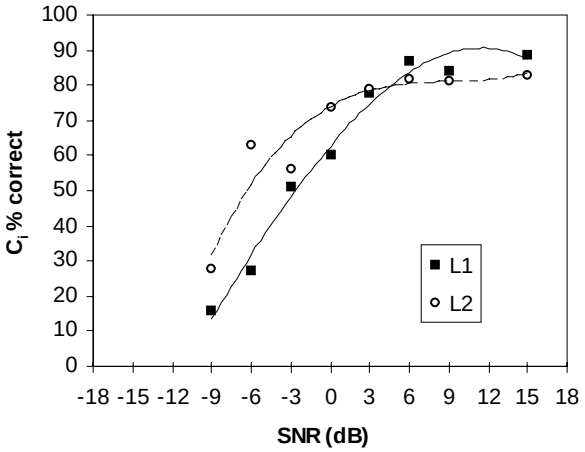

Figure 4. Initial consonant recognition score as a function of speech-to-noise ratio, for a single L1 talker (L1M4) and a single L2 talker (L2M7). Results are mean values for 4 L1 listeners. The lines are third order polynomial fits of the data.

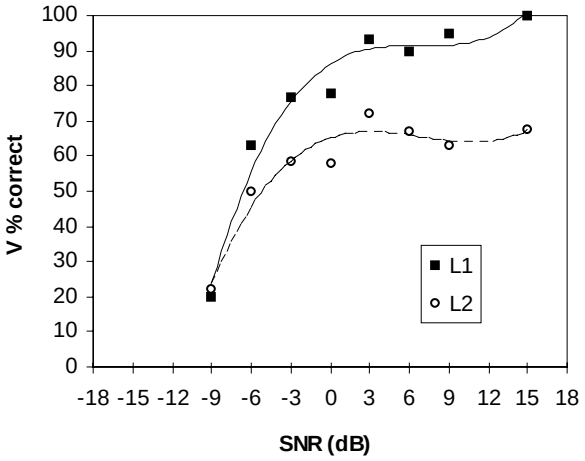

Figure 5. Vowel recognition score as a function of speech-to-noise ratio, for a single L1 talker (L1M4) and a single L2 talker (L2M7). Results are mean values for 4 L1 listeners. The lines are third order polynomial fits of the data.

This is more clearly the case with the vowels; for the L2-talker, vowel recognition saturates at a much lower percentage of correctly recognized vowels. This indicates that, irrespective of speech-to-noise ratio, some vowels by the L2 talker are consistently confused.

At two speech-to-noise ratios (-3 and +15 dB), phoneme recognition was measured for all 8 talkers, with 4 L1 and 4 L2 listeners. Results are shown in figures 6 and 7.

Figures 6 and 7 show, that differences between L1 and L2 speech intelligibility are cause mainly by the vowels. This is in agreement with the data presented in figures 4 and 5.

Non-nativeness of either talkers or listeners has a strong effect on vowel recognition, as may be verified by

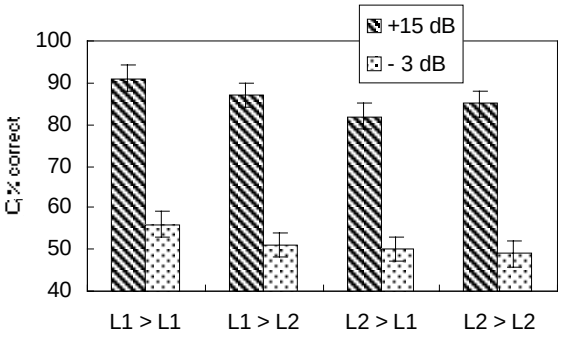

Figure 6. Initial consonant recognition scores at SNR-values of –3 and +15 dB. Results are averages (and standard errors) for 16 talker-listener pairs.

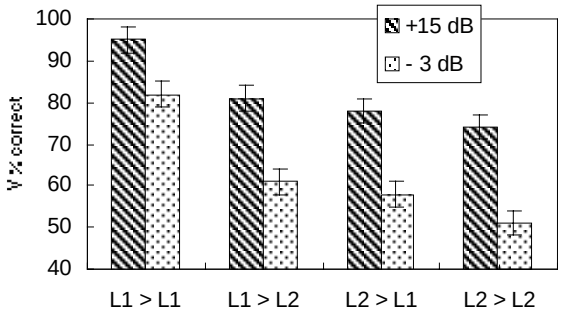

Figure 7. Vowel recognition scores at SNR-values of –3 and +15 dB. Results are averages (and standard errors) for 16 talker-listener pairs.

comparing the L1>L2 and the L2>L1 conditions on one hand, to the L1>L1 condition (baseline) on the other hand. In both cases (L2 talker or L2 listener) the difference in vowel recognition is around 15 percent-points in the +15 dB condition and more than 20 percent-points in the –3 dB condition. This suggests that the effect of additive noise on vowel recognition is somewhat stronger when non-natives are involved.

The loss of vowel intelligibility due to having a L2 talker, is not influenced much by also having a L2 listener. One might hypothesize that a L2 listener would be able to recognize and interpret the L2 accent better, hence recognizing vowels by L2 talkers more effectively. This is not the case, the L2>L1 scores are even slightly higher than the L2>L2 scores. This is consistent with the results from the SRT experiment.

4. ANALYSIS OF VOWEL CONFUSIONS

In order to perform a more diagnostic analysis of vowel confusions, confusion matrices were calculated from the phoneme responses. Although results were obtained at various SNR conditions, only the –3 and +15 dB results included all talkers. In order to obtain sufficiently 'filled' matrices, joint confusion matrices were calculated over both the –3 dB and +15 dB SNR conditions. This way, four matrices were obtained, corresponding to the four L1 and L2 talker-listener combinations. Each matrix contained 32 responses for each vowel (2 SNR

conditions, 4 talkers, 4 listeners). Unfortunately, the dataset was insufficiently large to perform meaningful multi-dimensional scaling analyses, which otherwise could have been used to construct 'nativeness-dependent' vowelspaces.

For each of the 15 vowels, in each condition, two types of confusion scores may be calculated from the confusion matrices: the percentage of *false positive* and the percentage of *false negative* responses. A false negative response is the failure to correctly respond with a phoneme upon presentation with that specific phoneme; a false positive response, is responding with that phoneme upon presentation of another phoneme.

The false negative scores are relatively robust, psychophysical indicators of phoneme recognizability; the paradigm is such, that a small false-negative error actually means good phoneme recognition in practice, and vice versa. The meaning of the false-positive error score is different; a large false-positive error may indicate consistent misarticulation of vowels in such a way that they all resemble another vowel; however, it may also reflect a measure of doubt of the listener. Even a vowel that is recognized fairly well as a stimulus, may attract false-positive responses as a response category. Such a response bias may occur, if listeners subjectively classify this vowel as 'difficult' and it as a response to any unrecognized (or similar-sounding) stimulus.

Of the 15 tested vowels, 8 were selected for further analysis. This set of 8 vowels comprised the 5 vowels with the highest overall false-positive scores, and the 5 vowels with the highest overall false-negative scores. The set consists of 6 monophtongs (/å/, /œ/, /y…/, /Š/, /o/, /Ø…/) and 2 diphtongs (/œy/, /åu/). Of this set of vowels, three are not normally found in American English: /y…/, /Ø…/ and /œy/. The 8 vowels within the set contribute 64% to the total number of false-negative responses, and 74% to the total number of false-positive responses of all 15 vowels. For the L1>L1 experiment, vowel recognition error scores are given in figure 8.

Note that the false-positive error rate is not limited to a maximum of 100%, since the number of times a vowel is "recognized" when it is not presented is only limited by the total number of vowel presentations.

All error scores in figure 8 are relatively low. The highest percentage of confusions occur with the vowel /o/.

In figures 9, 10 and 11, similar data is given as presented in figure 8, but now for the L2>L2, L1>L2 and L2>L1 experiments.

In figure 9, the distribution of false-negative responses over the vowels is quite different from the distribution of false-positive responses. Remarkably high false-positive scores are observed for the vowels /Ø…/ and /œy/, two of the vowels that do not occur in regular American English.

Figure 10 shows a closer correlation between false-positive and false-negative responses than figure 9, with the exception of the vowel /Ø…/.

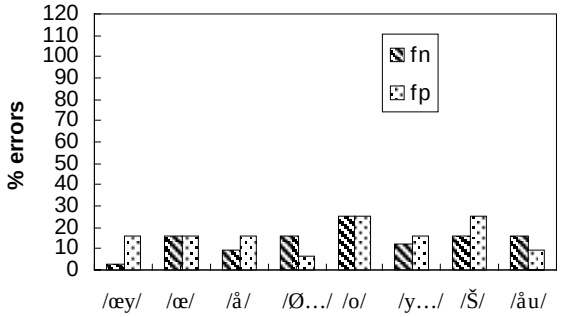

Figure 8. False-positive and false-negative responses in the L1>L1 experiment, to a limited set of vowels. An error score of 100% corresponds to 32 false responses

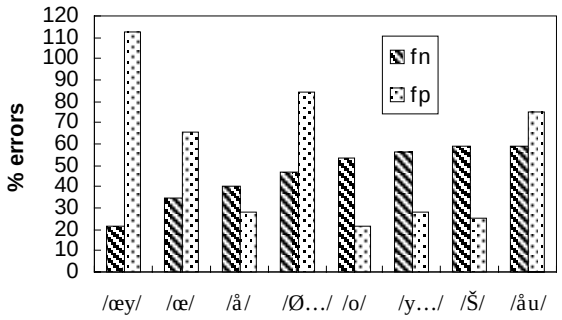

Figure 9. False-positive and false-negative responses in the L2>L2 experiment, to a limited set of vowels.

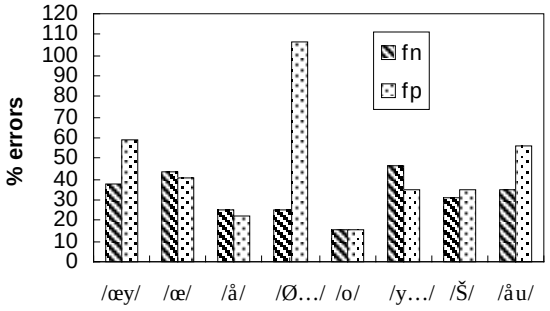

Figure 10. False-positive and false-negative responses in the L1>L2 experiment, to a limited set of vowels.

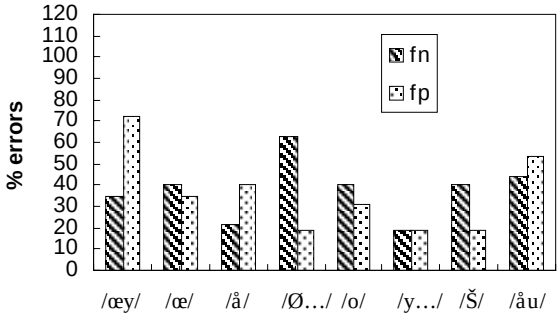

Figure 11. False-positive and false-negative responses in the L2>L1 experiment, to a limited set of vowels.

The vowel recognition errors are considered to be originating largely from two different error sources: non-nativeness of talkers, and non-nativeness of listeners. This is illustrated by the fact that the error scores in figure 8 (only native) are small in comparison to figures 9, 10 and 11.

The highest false-negative score in the L2>L1 experiment is obtained with the vowel /Ø…/; this indicates that unusual articulation of this non-English vowel by L2 talkers leads to reduced recognition by L1 listeners. Most of the other vowels also show higher error scores than in the L1>L1 experiment, which indicates that other vowels suffer from unusual articulation as well.

In the L1>L2 experiment, the highest false-negative score is of the non-English vowel /y…/, closely followed by several other vowels. Although the distribution of errors over vowels is somewhat different, the general tendency is similar to the L2>L1 case.
The largest false-positive scores in the L2>L1 experiment are /œy/ and /åu/. Many of these responses are given upon presentation of L2-versions of /Ø…/, which are usually very close to /œy/ or /åu/.
Two vowels, /Ø…/ and /œy/, lead to remarkably high false-positive recognition by L2 listeners (L1>L2 and L2>L2 experiments). Not many of the /Ø…/ and /œy/ presentations are *missed*, but at the expense of much false recognition. All this reflects the relatively poor model by the L2 listeners of the place of non-English vowels among other vowels.

5. CONCLUSIONS

Two types of speech intelligibility tests (SRT en CVC) produced results that correspond well. Both test types may be used to quantify the effect of non-nativeness on speech intelligibility. The advantage of the CVC test is the diagnostic value of the confusion matrices that may be generated.
Speech intelligibility of L2 (American) talkers of the Dutch language by Dutch listeners is less than L1 (native Dutch) speech intelligibility. The difference corresponds to approximately 3 dB difference in speech-to-noise ratio.
The main cause is consistent confusion of vowels, specifically those that do not occur in American English. This confusion is introduced by L2 talkers, but also by L2 listeners. The total degradation caused by introducing L2 talkers is slightly enhanced (certainly not reduced) by also having L2 listeners.

REFERENCES

[1] Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: the neighbourhood activiation model. *Ear & Hearing*, 19, pp. 1-36.

[2] Hood, J.D. & Poole, J.P. (1980). Influence of the speaker and other factors affecting speech intelligibility. *Audiology*, 19, pp 434-455.

[3] Bradlow, A.R., Toretta, G.M. & Pisoni, D.B. (1996). Intelligibility of normal speech I: global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20, pp 255-272.

[4] Sommers, M.S., Nygaard, L.C. & Pisoni, D.B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, 96(3), 1314-1324.

[5] Cox, R., Alexander, G.C. & Gilmore, C. (1987). Intelligibility of average talkers in typical listening environments. *Journal of the Acoustical Society of America*, 81(5), 1598-1608.

[6] Munro, M. J. (1999). The role of speaking rate in the perception of L2 speech. *Journal of the Acoustical Society of America*, 105(2), p. 1032.

[7] Cutler, A. (1999). Phonemic repertoire effects in lexical activation. *Journal of the Acoustical Society of America*, 105(2), p. 1033.

[8] Magen, H.S. (1998). The perception of foreign accented speech. *Journal of Phonetics*, 26, pp. 381-400.

[9] Bradlow, A.R. & Pisoni, D.B. (1998). Recognition of spoken words by native and non-native listeners: talker-, listener- and item-related factors. Research on spoken language processing , progress report No. 22, pp 74-94: Speech research laboratory, Department of Psychology, Indiana University.

[10] Ingram, J.C.L. & Park, S-G. (1998). Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners. *Journal of the Acoustical Society of America*, 103(2), 1161-1174.

[11] Flege, J.E., Bohn, O-S & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of phonetics*, 25, pp. 437-470.

[12] Plomp, R. & Mimpen, A.M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 18, pp. 43-52.

[13] Steeneken, H.J.M. (1992). On measuring and predicting speech intelligibility. Doctoral dissertation, University of Amsterdam.

[14] Miller, G.A. & Nicely, P. (1955). An analysis of perceptual confusion among some English consonants. *Journal of the Acoustical Society of America*, 27, 338-352.

[15] Pols, L.C.W. (1977). Spectral Analysis and identification of Dutch vowels in monosyllabic words. Doctoral dissertation, Free University of Amsterdam.

[16] Mayo, L.H., Florentine, W & Buus, S. (1997). Age of second-language acquisition and perception of Speech in noise. *Journal of Speech, Language and Hearing Research*, 40, 686-693.