ELSEVIER

Speech Communication 35 (2001) 103–113

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Intelligibility of native and non-native Dutch speech

Sander J. van Wijngaarden *

TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, Netherlands

Abstract

The intelligibility of speech is known to be lower if the speaker is non-native instead of native for the given language. This study is aimed at quantifying the overall degradation due to limitations of non-native speakers of Dutch, specifically of Dutch-speaking Americans who have lived in the Netherlands 1–3 years. Experiments were focused on phoneme intelligibility and sentence intelligibility, 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 speakers and American speakers of Dutch, using native Dutch listeners, was found to correspond to a difference in speech-to-noise ratio (SNR) of approximately 3 dB. The main segmental contribution to the degradation of speech intelligibility by introducing non-native speakers and/or listeners is the confusion of vowels, especially those that do not occur in American English. Vowels that are difficult for second-language speakers to produce are also difficult for second-language listeners to classify; such vowels attract false recognition, reducing the overall recognition rate for all vowels.

1. Introduction

Even in the Information Age, speech as a means of communication between humans is as important as ever before. At the same time, because physical distances are not as important anymore, the effectiveness of human speech communication increasingly suffers from language barriers. In a telephone conversation between two people at opposite sides of the world, at least one of both is likely to be speaking in a language other than his or her native tongue.

Although second-language (L2) speakers may be fluent in a given language, we usually subjectively find their speech less intelligible than native speech. The reverse is also true: non-native listeners have more difficulty understanding speech than native listeners. For many practical applications, it is important to have some quantitative description of the loss of intelligibility when non-native speakers or listeners are involved. This knowledge may help to design systems with enough intelligibility 'headroom' to cope with non-native users. Some examples could be: public address systems at airports, radio communications in multi-lingual military forces, or compressed-speech internet news broadcasts.

Unfortunately, speech-degrading influences tend to affect non-natives stronger than natives. This has for instance been proven, in the case of non-native *listeners*, for noise (Bergman, 1980; Gat and Keith, 1978; Florentine, 1985) and reverberation (Nábělek and Donahue, 1984).

The reasons why non-natives are more severely affected by speech degradations are considered

[*] Tel.: +31-346356230.

E-mail address: vanwijngaarden@tm.tno.nl (S.J. van Wijngaarden).

0167-6393/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(00)00098-4

to be at multiple levels. Florentine (1985) interprets the shallower psychometric curves found in sentence intelligibility tests as a lack of effective redundancy. A more limited knowledge of phonological, lexical, syntactic and semantic constraints leads to a reduced use of the inherent redundancy in test sentences. Bradlow and Pisoni (1999) state that non-native listeners have difficulty with fine phonetic discrimination at the segmental level. They base this conclusion on a strong dissociation between word recognition of lexically 'easy' and 'hard' words, even when controlled for word familiarity.

The existence of a strong relation between the amount of L2 experience and speech intelligibility is well established, both for L2 speakers and L2 listeners (for instance: Buus et al., 1986; Flege et al., 1997). Also, the age of L2 acquisition is important (Mayo et al., 1997). When studying general effects of non-nativeness on intelligibility, these parameters should be controlled.

The main object of this research is to quantify the effect of non-nativeness, both of speakers and listeners, on the effectiveness of human speech communication. A secondary object is to gain insight into the intelligibility-degrading factors associated with non-native *speakers*. When listening to non-native speakers, one can often immediately identify two factors that may reduce intelligibility: speech sounds are produced in an unusual, unexpected way ('distorted' phoneme inventory), and sentences are intoned in an unusual fashion. An attempt is made to assess the influence of these two factors separately.

These objectives may be achieved by carrying out speech intelligibility experiments with L1 (native) and L2 (non-native) subjects (speakers/listeners) in a certain language, in our case Dutch. The study is limited to fluent non-native speakers whose first language is American English. Since the effects of non-nativeness on intelligibility are expected to be especially apparent with degraded speech, intelligibility experiments are carried out with speech degraded by additive noise. A native Dutch control group is included in all experiments.

2. Experimental setup and methods

2.1. Intelligibility test types

A choice has to be made for which speech-unit level to test at: at the sentence, word or phoneme level. As the influence of non-nativeness is expected at both segmental *and* supra-segmental levels, both a sentence and a word/phoneme test is to be used.

With sentence intelligibility tests, the influence of redundancy introduced by linguistic constraints, such as semantic and syntactic constraints, is included in the outcome. Some supra-segmental *acoustic* speech features that may contribute to the overall redundancy of speech are also tested. Since unusual intonation of a sentence may reduce predictability, factors such as pitch contours, temporal envelope and syllabic rhythm potentially have an effect on sentence intelligibility. When using a sentence intelligibility test, these features may even be manipulated to gain insight into the importance of each feature to the overall intelligibility of speech.

Phoneme recognition tests are not (or hardly) sensitive to supra-segmental, redundancy-related factors. Since it seems reasonable to assume a major influence of non-nativeness on segmental phonetic-acoustics, phoneme tests serve to give relatively detailed information at this level.

A complicating factor in carrying out the experiments is that the experimental paradigm should 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 the typing out of nonsense words by L2 listeners, who will have a tendency to use native-language spelling of some nonsense words).

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 (Plomp and Mimpen, 1979). The phoneme intelligibility test is closely related to the equally balanced CVC test (Steeneken, 1992).

2.2. Speech reception threshold method

The SRT method gives a robust measure for sentence intelligibility in noise, corresponding to the speech-to-noise ratio (SNR) that gives 50% correct response for short, redundant everyday sentences.

In the SRT testing procedure, masking noise is added to test sentences in order to obtain the required SNR. The masking noise spectrum is equal to the long-term spectrum of the test sentences. After presentation of each sentence, a subject responds by repeating 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 to correctly, using 4 dB steps. This is done to quickly converge to the 50% intelligibility threshold. By taking the average SNR 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 (cp. Miller and Nicely, 1955).The method is similar to an open-response equally balanced CVC paradigm (Steeneken, 1992). The main differences are that the final consonant is not tested, and that the subject responds by choosing an alternative from a (nearly) exhaustive list of possible CVC words, instead of typing the word in response to the stimulus. The advantage of this approach is that extensive training of subjects becomes unnecessary, while the construction of confusion matrices is still possible. Problems that were expected using a 'difficult' open-response paradigm with non-native subjects were successfully avoided.

During each 3–4 min 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 a computer display, and for a vowel 15 alternatives. When testing the vowel /øː/, for instance, the list of CVC words for the listener to choose from could be 'jaap', 'jup', 'jeup', 'jip', etc.; between the alternatives to choose from, the only difference is the vowel (rhyme word concept).

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

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 digitised 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 speakers of the Dutch language, matching all criteria, are rather difficult to find, we chose to record the material at a location of the speaker's choice. This proved to be an effective measure to facilitate the recruitment of subjects, but led 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 effect 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.

After collection of the SRT sentences, it was verified that at a very benign SNR (+15 dB) the intelligibility of the sentences was (nearly) 100% for all L1 and L2 speakers, when presented to L1 listeners. This was done to make sure that the intelligibility of the 'clear' speech was not too close to 50% sentence intelligibility, in which case the validity of SRT results would have been at stake.

2.5. Subjects

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

As stated in the introduction, perception and production of foreign speech sounds depends on the experience that subjects have in a foreign language, while 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 (Flege et al., 1997; Mayo et al., 1997). Three of the four L2 speakers 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 who might be categorised as 'early bilingual', showed appreciably better control of the Dutch language, the American accent was still readily noticed.

The L1 speakers were selected to match the L2 speaker 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 at 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. Experimental results

3.1. Sentence intelligibility

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

The results given in Fig. 1 were analysed statistically by performing an analysis of variance (ANOVA). Very significant main effects ($p < 0.01$)

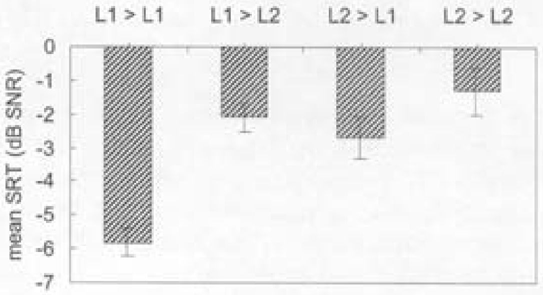

Fig. 1. Results for four types of speaker–listener combinations (16 speaker–listener pairs per condition, mean values and standard errors given). L1 > L2, for instance, means native speaker, non-native listener.

were found for speaker native language, listener native language, and individual speaker. Differences between the four categories in Fig. 1 were evaluated by means of planned comparisons. The difference in SRT results between L1 > L2 and L2 > L2 is not significant; all other differences in Fig. 1, although some are quite small, are statistically significant ($p < 0.05$).

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

The condition with L1 speakers and L2 listeners requires a nearly 4 dB lower noise level for the same 50% sentence intelligibility than the L1 > L1 condition. The L2 > L1 condition (L2 speakers, L1 listeners) also allows less noise for 50% sentence intelligibility; the difference is now 3 dB. The L2 > L2 condition, showing the lowest intelligibility results, allows for 4.5 dB less noise (although the difference with L1 > L2 is not significant).

Fig. 1 gives a general overview of the influence of non-nativeness of speakers and listeners on speech intelligibility, at least for the particular languages involved (Dutch and American English). It also shows that, even though the L2 speaker group was less experienced than the L2 listener group, having L2 listeners gives relatively slightly more degradation of speech intelligibility than having L2 speakers ($p < 0.05$). The combination of L2 listeners *and* L2 speakers gives an additional degradation that is less than the degradation caused by L2 speakers and L2 listeners separately.

The results of Fig. 1 are also given in Figs. 2 and 3, but now by individual speaker instead of speaker/listener group. For the L1 listener group (Fig. 2), all L1 speakers appear to offer better intelligibility than any L2 speaker, although not all differences are statistically significant (for example, speaker L2F8 is not significantly different from L1M1, L1F4 and L2M7). However, each individual L2 speaker *is* significantly different from the L1 speakers *as a group*, and vice versa.

Fig. 3 (same as Fig. 2, but now for L2 listeners) shows somewhat different results; to L2 listeners, the highest intelligibility is offered by one of the L2 speakers (L2M7). The individual L2 speakers L2M7 and L2F8 are *not* significantly different from the L1 speakers as a group. This may seem surprising, since the average score by L2 speakers as shown in Fig. 1 is quite low. However, this is mainly because of speakers L2M5 and L2F6; these same two speakers showed the lowest intelligibility with L1 listeners. Upon being asked about perceived accents, L1 listeners indicated that these two speakers had the strongest accents. In particular, remarks were made about the unusual intonation of sentences. Less authentic intonation by L2 speakers may partially explain results in both Fig. 2 (native listeners) and Fig. 3 (non-native listeners). Non-native intonation may contribute less to the overall redundancy in the SRT sentences.

To investigate this hypothesis, the SRT experiments were repeated with a different native Dutch

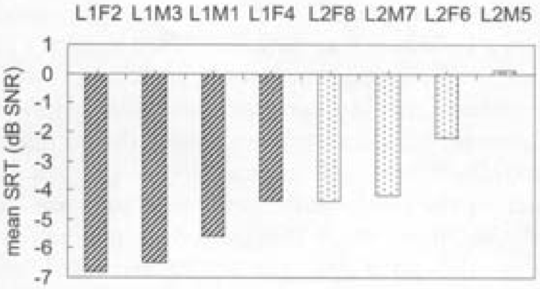

Fig. 2. Mean SRT scores for eight individual speakers, within the L1 group of listeners (four listeners per condition). L2M5, for instance, means L2 talker, male, talker #5.

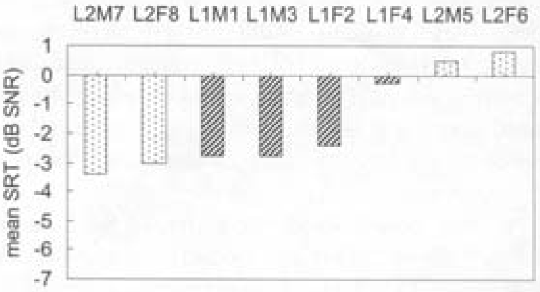

Fig. 3. Mean SRT scores for eight individual speakers, within the L2 group of listeners (four listeners per condition).

listener group. Besides the 'standard' SRT condition, a condition was now also tested in which the speakers' voice pitch was made uniform, in order to reduce the intonation. Pitch contours were measured for each sentence, using the computer program "praat" (Boersma, 1999). The average pitch of the voiced parts in the utterance was calculated, after which the pitch was manipulated using the pitch-synchronous overlap and add (PSOLA) algorithm, setting the pitch of all voiced parts to the average of the sentence. This procedure, besides effectively making the sentences more monotonous, introduced minor artifacts in the speech signal. In practice, since in the SRT procedure speech is always mixed with noise at a low SNR, these artifacts could not be distinguished during testing. None of the subjects, who were explicitly asked for this, could confirm having heard anything unusual about the manipulated speech, apart from the very monotonous character of the speech.

As is to be expected, the standard SRT results in Fig. 4 do not differ significantly from the results as given in Fig. 1 (same experiment, different listeners). The sentences with uniform pitch were less intelligible than the standard sentences. For native speakers the difference is 2 dB, for non-native speakers 3 dB. Although this effect may partly be caused by speech degradation due to the PSOLA manipulations, it is likely that the partial loss of intonation by removing pitch cues causes a reduction of the intelligibility. Intonation contributes to the redundancy in a sentence; reducing intonation makes speech less resistant to noise. The difference in the effect between native and non-native speakers, although statistically significant, is relatively small (1 dB); this indicates that the contribution of intonation to overall intelligibility is roughly the same for native and non-native speakers.

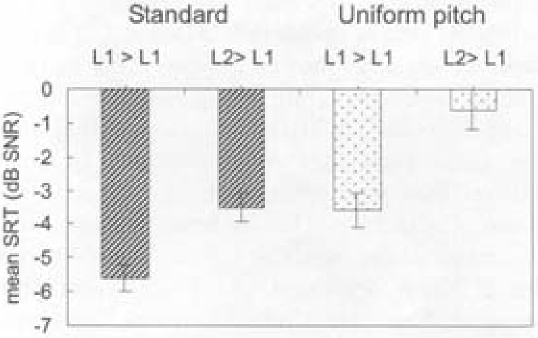

Fig. 4. SRT results with a new group of native listeners (16 speaker–listener pairs per condition, mean values and standard errors given). Both the standard SRT condition and a condition with uniform pitch were tested.

We hypothesised that the unusual intonation in non-native speech may carry less information than intonation in native speech. That hypothesis does not seem as likely anymore, given the small difference between L1 and L2 speakers in the effect of pitch manipulation. Furthermore, as shown in Fig. 5, the correlation between standard SRT and SRT with modified pith is relatively high (correlation coefficient $R = 0.93$ across eight speakers). This means that the effect of pitch manipulation on intelligibility is roughly the same for all speakers, and is not very dependent on individual speaker characteristics and non-native accent.

3.2. Phoneme intelligibility

When measuring sentence intelligibility, effects at both the segmental and the supra-segmental levels are included in the measurement results. By

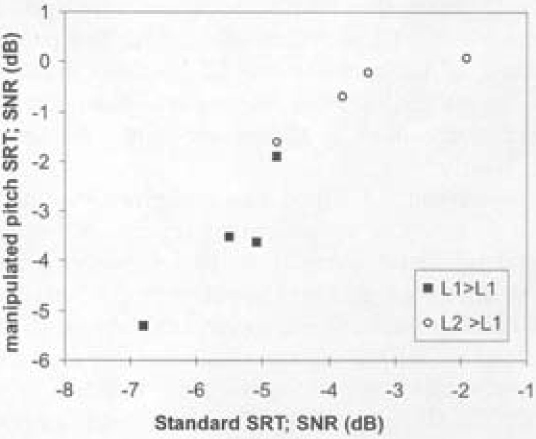

Fig. 5. Correlation between standard SRT and SRT with manipulated pitch ($R = 0.93$), across four native and four non-native speakers (mean value of the same four listeners for every data point).

contrast, phoneme intelligibility tests identify effects at the segmental level only. CVC experiments were performed at various SNRs, for a single native and a single non-native speaker (four native listeners). Of the four non-native speakers, the speaker was chosen whose accent was most easily recognised as American English by the L1 listeners. Results for this CVC experiment, for initial consonants and vowels separately, are given in Figs. 6 and 7.

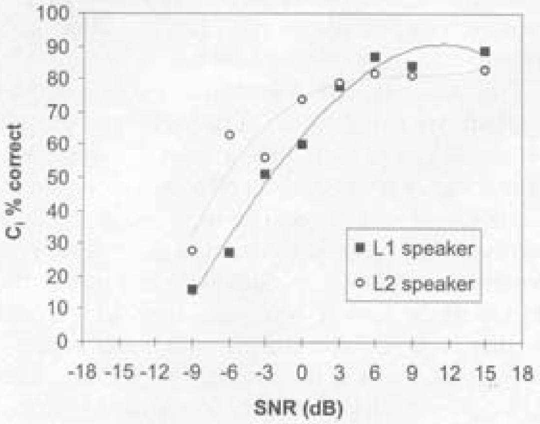

Fig. 6. Initial consonant recognition score as a function of SNR, for a single L1 speaker (L1M4) and a single L2 speaker (L2M7). Results are mean values for four L1 listeners. To guide the eye, third-order polynomial fits are drawn.

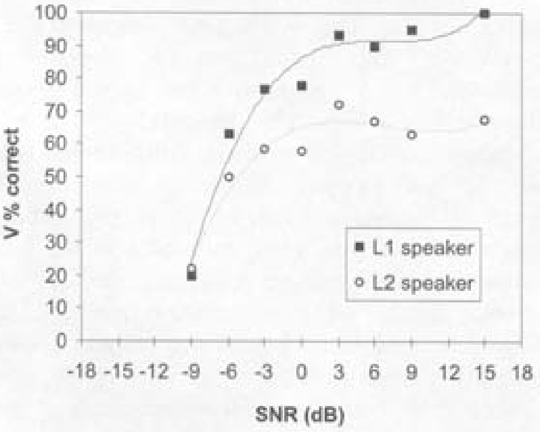

Fig. 7. Vowel recognition score as a function of SNR, for a single L1 speaker (L1M4) and a single L2 speaker (L2M7). Results are mean values for four L1 listeners. To guide the eye, third-order polynomial fits are drawn.

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.

For the initial consonant recognition, no clear differences between L1 and L2 speakers are observed. For the vowels, there is a clear difference between the L1 and L2 speaker; for the L2 speaker, vowel recognition saturates at a much lower percentage of correctly recognised vowels (around 60%). This indicates that, irrespective of SNR, some vowels by the L2 speaker are consistently confused.

At two SNRs (−3 and +15 dB), phoneme recognition was measured for all eight speakers, with four L1 and four L2 listeners. Results are shown in Figs. 8 and 9.

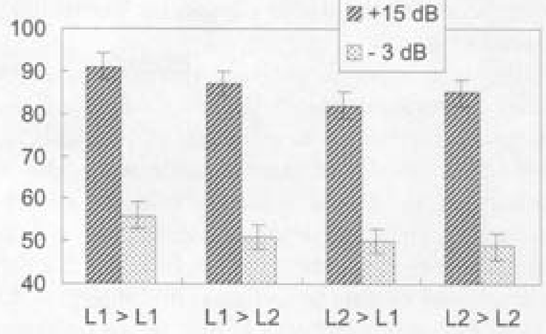

Fig. 8. Initial consonant recognition scores at SNR values of −3 and +15 dB. Results are averages (and standard errors) for 16 speaker–listener pairs.

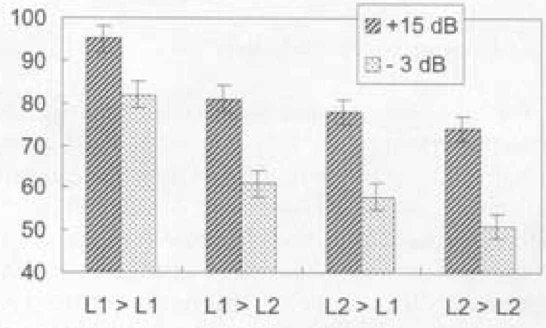

Fig. 9. Vowel recognition scores at SNR values of −3 and +15 dB. Results are averages (and standard errors) for 16 speaker–listener pairs.

Figs. 8 and 9 show that differences between L1 and L2 speech intelligibility in terms of phoneme confusions are caused mainly by the vowels; vowels show a much stronger effect of non-nativeness of speakers *and* listeners than initial consonants. This is observed by comparing the L1 > L2 and the L2 > L1 conditions on one hand, to the L1 > L1 condition (baseline) on the other hand. The difference in vowel recognition is 17–24% ($p < 0.01$), while the difference in consonant recognition is only 4–9%. Within the consonant recognition scores, the only significant difference is the 9% difference between L1 > L1 and L2 > L1 (SNR +15 dB).

On average, the SNR has a clear effect on recognition of vowels by L2 speakers. The difference between an SNR of +15 and −3 dB is 20%. The saturation effect observed in Fig. 7 for a single speaker appears to be at a higher average recognition score for the whole L2 speaker group (80% instead of 60%).

It is interesting to see that there seems to be some correspondence between *production* and *perception* by non-natives; the loss of intelligibility seems to be nearly the same in both cases. Also in both cases, the largest effects are in the vowels. One might therefore hypothesise that in a fully non-native configuration (L2 > L2), the L2 listeners would be able to recognise and interpret the typical L2 speech patterns better, hence recognising speech by L2 speakers more effectively. This is not the case; both for CVC and SRT the L2 > L1 scores are the same or even slightly higher than the L2 > L2 scores.

4. Analysis of vowel confusions

We have observed that vowel confusions seem to be an important factor for non-native speech intelligibility. 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 speakers. To avoid having matrices that are too sparse for meaningful analysis, 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 speaker–listener combinations. Each matrix contained 32 responses for each vowel (two SNR conditions, four speakers, four listeners).

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 negative* and the percentage of *false positive* responses. A false negative response is the failure to correctly respond with a phoneme upon presentation of 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 recognition; 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 recognised 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 use it as a response to any unrecognised (or similar-sounding) stimulus.

Of the 15 tested vowels, 8 were selected for further analysis. This 8-vowel set 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 monophthongs (/ɑ œ yː ɔ o øː/ and 2 diphthongs (/œy ɑu/). Of these vowels, three are not normally found in American English: /yː øː 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 Fig. 10.

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

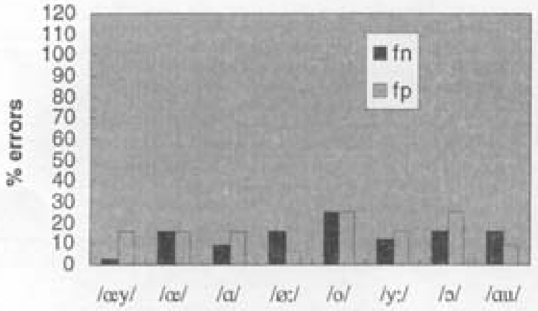

Fig. 10. False-positive and false-negative responses in the L1 > L1 experiment, to a limited set of vowels. An error score of 100% corresponds to 32 false responses.

All error scores in Fig. 10 are relatively low. The highest percentage of confusions occur with the vowel /o/. In Figs. 11–13, similar data are given as presented in Fig. 10, but now for the L2 > L2, L1 > L2 and L2 > L1 experiments, respectively.

In Fig. 11, 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.

The highest false-negative score in the L2 > L1 experiment (Fig. 13) is obtained for the vowel /ø:/; this indicates that unusual articulation of this non-English vowel by L2 speakers leads to reduced recognition by L1 listeners. Interestingly, it is for this same vowel that the highest *false-positive* error is found in the reverse (L1 > L2) situation (Fig. 12). This suggests that there is an awareness among L2 subjects that their control of this vowel is poor, which apparently leads to a form of hypercorrection – a high false recognition rate for difficult vowels.

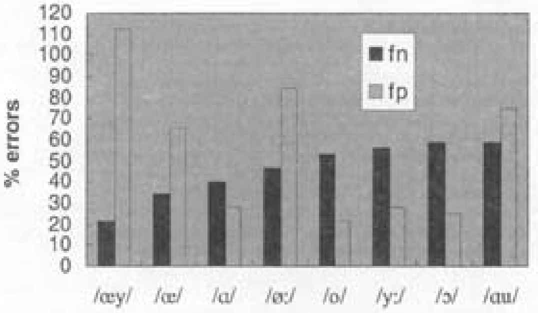

Fig. 11. False-positive and false-negative responses in the L2 > L2 experiment, to a limited set of vowels.

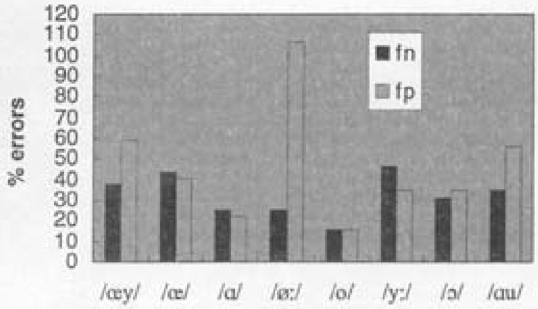

Fig. 12. False-positive and false-negative responses in the L1 > L2 experiment, to a limited set of vowels.

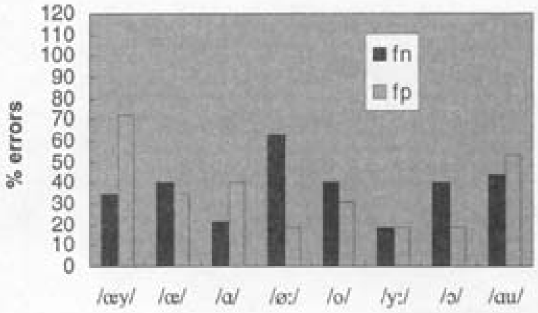

Fig. 13. False-positive and false-negative responses in the L2 > L1 experiment, to a limited set of vowels.

To investigate the link between non-native vowel production and perception, the correlations between the L1 > L2 and L2 > L1 recognition errors are calculated. The correlation coefficients are given in Table 1.

In Table 1, the L1 > L1 error scores are included to check for correlations stemming from factors other than non-nativeness. Such correlations appear not to be present. Significant (and positive) correlations are only found between: L1 > L2 (fp) with L2 > L1 (fn), and L2 > L1 (fp) with L1 > L2 (fn).

Let us assume that the contribution to the overall error scores caused by L1 listeners and speakers are negligible, and interpret what these correlations imply for L2 production and perception. The first of the two significant correlations then suggests (as before) that vowels that

Table 1
Correlations between vowel recognition error scores, across all 15 tested vowels

		False negative error			False positive error		
		L1 > L1	L1 > L2	L2 > L1	L1 > L1	L1 > L2	L2 > L1
False negative error	L1 > L1	–	−0.04	0.36	0.06	0.29	0.25
	L1 > L2	−0.04	–	0.28	0.27	0.21	0.77[a]
	L2 > L1	0.36	0.28	–	−0.23	0.62[a]	0.22
False positive error	L1 > L1	0.06	0.27	−0.23	–	0.09	0.22
	L1 > L2	0.29	0.21	0.62[a]	0.09	–	0.24
	L2 > L1	0.25	0.77[a]	0.22	0.22	0.24	–

[a] Significant correlation coefficients ($p < 0.05$).

are difficult to produce tend to be perceived, even when *not* presented. The second correlation suggests that vowels that are difficult for non-natives to recognise tend to be produced by non-natives in such a way that they sound like other vowels. Production and perception appear to be influenced by the imperfections of a joint (personal) model of a non-native vowel space.

5. Conclusions

Two types of speech intelligibility tests (SRT and CVC) produced results that indicate an appreciable effect of non-nativeness on speech intelligibility. The SRT results are the easiest to apply to speech communication in practice: for speech communication with (fluent) non-natives in the presence of noise, a 3–4 dB better SNR is required to reach the same speech intelligibility as in fully native speech communication.

CVC results indicate that there is an appreciable effect of non-nativeness on phoneme recognition. The recognition of vowels is affected more than recognition of consonants. This is partly caused by consistent (SNR-independent) confusion of vowels, specifically those vowels that do not occur in American English.

Sentence intelligibility, as measured with the SRT method, is partly a measure of the redundancy in spoken messages. From the experiments described, no conclusions can be drawn regarding how important differences in effective redundancy between L1 and L2 subjects are; the SRT test also incorporates other factors. It is clear from the CVC experiments, however, that the loss of intelligibility is at least partially at the phoneme level. Less authentic intonation of sentences, as occurs with some L2 speakers, was not found to be an important cause of decreased intelligibility.

The errors in L2 production and L2 perception of vowels appear to be connected. Vowels that are difficult to produce are also difficult to classify; such vowels attract false recognition, reducing the overall recognition rate for all vowels.

References

Bergman, M., 1980. Aging and the Perception of Speech. University Park Press, Baltimore.

Boersma, P., 1999. Praat 3.8.12; A system for doing phonetics by computer. University of Amsterdam, Amsterdam.

Bradlow, A.R., Pisoni, D.B., 1999. Recognition of spoken words by native and non-native listeners: talker-, listener- and item-related factors. Journal of the Acoustic Society of America 106 (4), 2074–2085.

Buus, S., Florentine, M., Scharf, B., Canevet, G., 1986. Native, French listeners' perception of American-English in noise. In: Proceedings of Internoise 86, pp. 895–898.

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, 437–470.

Florentine, M., 1985. Non-native listeners' perception of American-English in noise. In: Proceedings of Internoise 85, pp. 1021–1024.

Gat, I.W., Keith, R.W., 1978. An effect of linguistic experience; auditory word discrimination by native and non-native speakers of English. Audiology 17, 339–345.

Mayo, L.H., Florentine, M., Buus, S., 1997. Age of second-language acquisition and perception of speech in noise. Journal of Speech, Language and Hearing Research 40, 686–693.

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.

Nábělek, A.K., Donahue, A.M., 1984. Perception of consonants in reverberation by native and non-native listeners. Journal of the Acoustical Society of America 75 (2), 632–634.

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

Pols, L.C.W., 1977. Spectral analysis and identification of Dutch vowels in monosyllabic words. Doctoral Dissertation, Free University of Amsterdam.

Steeneken, H.J.M., 1992. On measuring and predicting speech intelligibility. Doctoral Dissertation, University of Amsterdam.