Intonation and the perceptual separation of simultaneous voices

Journal of Phonetics (1982) 10, 23–36

Intonation and the perceptual separation of simultaneous voices

J.P.L. Brokx
Dr. Neher Laboratorium, Leidschendam, The Netherlands

S.G. Nooteboom
Institute for Perception Research, Eindhoven, and Department of General Linguistics, Leiden University, The Netherlands

Received 12th February 1981

Abstract: The present paper examines the role of speech pitch in the perceptual separation of simultaneous speech messages, when both messages are spoken by the same speaker and there are no differences in directional hearing. In a first experiment, employing resynthesized speech with completely monotonous pitch, it is shown that intelligibility of the target message can be manipulated by introducing an artificial constant difference in pitch between target speech and interfering speech. Within certain limits intelligibility increases with increasing difference in pitch. In a second experiment natural speech is employed for both target and interfering messages. The interfering speech is always spoken with normal intonation, whereas the target messages are either spoken with normal intonation or deliberately spoken in a monotone. For both intonation conditions the messages are either spoken within the same pitch range as the interfering speech (SAME PITCH), or within a considerably higher pitch range (DIFFERENT PITCH). For the messages spoken with normal intonation the SAME PITCH condition is considerably less intelligible than the DIFFERENT PITCH condition. For the monotonously spoken messages the results are less clear. Here the effect of a difference in pitch range is probably confounded with the effects of other properties of speech which result from a monotonous pronunciation. The main results of these experiments can be related to the phenomenon of "perceptual fusion", occurring whenever two simultaneous sounds have identical pitches, and to "perceptual tracking": whenever the pitches of target and interfering speech cross each other, the listener runs the risk of inadvertently switching his attention from the target speech to the interfering speech.

Introduction

Over the past fifteen years a number of investigations have been directed at the description of the basic regularities inherent in Dutch intonation. This series of studies was initiated by Cohen & 't Hart (1967). It can be distinguished from many earlier intonational studies, particularly within the realm of linguistics, by a constant and deliberate attempt to form a non-functional description of perceptually relevant aspects of pitch contours, seen as the carriers of intonation patterns. The method of perceptual analysis-by-synthesis, replacing

0095–4470/82/010026+14 $02.00/0

the more traditional use of linguistic distinctiveness and described in detail by Cohen & 't Hart (1967) and 't Hart & Cohen (1973), has proved fruitful. It has led to an explicit set of rules generating stylized versions of pitch contours capturing the basic properties of Dutch intonation from a listener's point of view and applicable in the synthesis-by-rule of connected speech ('t Hart & Cohen, 1973; Collier, 1972; 't Hart & Collier, 1975).

Given such a basic description, work is now proceeding to discover the potential functions, linguistic or otherwise, of pitch contours in speech communication. The present paper results from one of several research projects focusing on various communicative functions of intonation. Others have been directed at the relation between pitch contours and accentuation (Van Katwijk, 1974), the interdependencies between pitch contours and syntactic structuring (Collier & 't Hart, 1975; De Rooij, 1979), the lexical, syntactic, and semantic sources of pitch accents, and the relation between pitch accents and comprehension (in progress). Attention has also been paid to the contribution of pitch contours to the perceived "auditory coherence" of speech (Nooteboom *et al.*, 1978; Brokx, 1979). By "auditory coherence" we mean the phenomenon that speech is normally perceived as coming from a single speaker, preserving a perceptual continuity which holds the speech together, and, as it were, continuously signals to the listener that he is still listening to the same message. Part of the experiments reported in Brokx (1979) were specifically concerned with the relation between discontinuities in the pitch contour and auditory incoherence in sequences of speech sounds. The main results of these experiments were reported in Nooteboom *et al.* (1978). In the present paper we re-examine some data from a second set of experiments described by Brokx (1979), focusing on the contribution of overall pitch and pitch contours to the perceptual separation of simultaneous voices.

At the time we observed that "auditory coherence" in the earlier mentioned sense is not a necessary precondition for speech intelligibility. If we listen to a stretch of speech in a quiet room, the speech may still be intelligible even if it contains many severe and unnatural discontinuities, for example obtained by artificially switching between a male and a female voice after each syllable. But it occurred to us that auditory coherence, and particularly the continuity in its main auditory correlate, the pitch contour, may become essential when one listens to one voice in the presence of other voices. In this way our interest was aroused in the role of pitch in the perceptual separation of simultaneous voices.

In exploring this further we have taken a cue from Cherry (1953) who addressed himself to the question of how one recognizes what one person is saying when others are speaking at the same time. Cherry mentioned as possible facilitating factors directional hearing, visual information, individual differences in voice characteristics and dialect, and transitional probability. Although his main experiments were directed at directional hearing and transitional probability, he also observed that, when all the above-mentioned factors except transitional probability were eliminated, by recording two messages spoken by the same speaker on the same magnetic tape, the result may sound "like a babel", but the messages can still be separated.

Of course, under these conditions transitional probability, the importance of which was convincingly demonstrated by Cherry, still operates. Directional hearing, which was also shown to be extremely helpful in separating simultaneous messages, is excluded, just as visual information is, and, because the same speaker is used for both interfering and target speech, individual differences in voice characteristics and dialect are also excluded. The contribution of these latter differences, particularly the difference between a male and a female voice, was demonstrated by Treisman (1964).

Egan *et al.* (1954) have shown that, when speech is masked by speech from the same

speaker, some measure of helpful individual differences can be artificially reintroduced by band-pass filtering, or by changing the intensity of either the interfering or the target speech. Band-pass filtering had a positive effect on intelligibility. In a similar vein the present experiments examine the intelligibility of speech interfered with by speech from the same speaker, while reintroducing some difference which normally exists between speech from different speakers, in this case differences in overall level of pitch and in pitch contour. The experiments by Treisman (1964) showing a considerable positive effect of the male–female difference on the intelligibility of the target speech, suggest that overall pitch level may be helpful. An experiment by Darwin (1975) and our own experiments (Nooteboom *et al.*, 1978) indicate that continuity of the pitch contour helps the listener to track one voice in the presence of another voice, or, if we think of the experiments to be described below, in the presence of the same voice speaking a second and interfering message.

Below we will describe two experiments examining the role of pitch differences in the perceptual separation of simultaneous voices. In Experiment I we have explored the possibility of manipulating the intelligibility of target speech utterances in the presence of interfering speech by artificially introducing a constant difference in pitch between target and interfering speech. As noted by Stumpf (1890), two simultaneous sounds having identical pitches tend to fuse perceptually, into a single sound, losing the characteristic qualities of the original sounds, whereas two simultaneous sounds differing in pitch may be perceptually separated and separately recognized. This fusing power of identical pitches has also been demonstrated for dichotically presented speech-like sounds (Broadbent & Ladefoged, 1957; Myers *et al.*, 1975). From the power of pitch to either fuse or separate two simultaneous sounds we predict that pitch separation enhances the intelligibility of speech interfered with by other speech. In order to gain precise control over the pitch of target and interfering speech, in Experiment I vocoderized speech is employed. In Experiment II we have used real speech, for both target and interfering messages. In doing so, we narrow the gap between the laboratory situation being investigated and those aspects of everyday-life speech communication we hope to illuminate by our findings. At the same time, of course, we lose precise control over the experimental variable we are most interested in, the difference in pitch between the two speech messages. For example, by varying the difference in overall level of pitch between two messages with normal intonation, we change not only the difference in pitch averaged over time, but also the frequency with which the two pitches cross each other. In an attempt to examine the effect of crossing pitches, we have not only used test utterances spoken with normal intonation, but also test utterances which were deliberately spoken in a monotone.

In both experiments reported below, the first employing resynthesized speech and the second natural speech, we have introduced the intensity of the target speech relative to the interfering speech as an experimental variable additional to those motivated above. This was done to guard against possible floor or ceiling effects.

Experiment I

This first experiment was set up to examine the effect of a constant difference in pitch between two speech messages on the intelligibility of one of these messages and to see whether intelligibility can be manipulated by varying the magnitude of this difference.

Method

The interfering speech was obtained as follows. A short story, approximately 600 words in length, was read aloud by a male speaker of Dutch who was instructed to speak with normal

intonation and to minimize speech pauses. His speech was recorded on disk in 8-bit words with a sampling frequency of 10 kHz, after low-pass filtering with a cutoff frequency of 4800 Hz. After removal of all remaining speech pauses, the digitized speech was subjected to LPC analysis and a subsequent formant analysis and then resynthesized with a Rockland digital speech synthesizer. F_0 was fixed at 100 Hz during all voiced portions of speech. The resynthesized speech was recorded on one track of a magnetic tape with a two-channel A 77 Revox tape recorder, and repeated as many times as necessary. This continuous stream of speech having no speech pauses was used as the interfering speech.

The test speech, whose intelligibility was to be measured, was obtained by having the same speaker read aloud short Dutch sentences. These sentences were of the type described for English by Nakatani & Dukes (1973), all sentences having the same syntactic frame, in this case article + substantive + verb + adverb + preposition + article + substantive. Each position was filled with a monosyllabic word. For each of the four content word positions (substantive, verb, adverb, substantive) the actual words were found by drawing at random with replacement from a set of common Dutch words. The sentences generated in this way were always syntactically correct and generally semantically anomalous. All semantically normal sentences were removed from the set, so that all test sentences were of the type exemplified by *the town swims now in a sheep*. Ninety-six of such sentences were used in this experiment. The resulting speech utterances were treated in the same way as the interfering speech (but no pauses, if any, were removed from these utterances), except that in resynthesizing F_0 was fixed at 100, 103, 106, 109, 120, or 200 Hz. Before recording the maximum intensities of these utterances were kept within a range of 2 dBA with the help of a Bruel & Kjaer dBA meter.

The test utterances were recorded on the second track of the magnetic tape in four blocks of 24 utterances, each block containing four utterances of each of the six F_0 levels, in a random order of F_0 levels. During the experiment the intensity ratio between test utterances and interfering speech was set before each of the four blocks of utterances. The intensity ratios were 0, −5, −10, and −15 dB. Because no objective measure could be found for the 0 dB intensity ratio to correspond to an impression of equal loudness of interfering speech and test utterances, probably because of differences between continuous speech and short utterances in their intensity fluctuations, the 0 dB ratio was defined as the relative intensity at which interfering speech and test utterances gave the impression of having roughly the same loudness. The other intensity ratios were found by simply reducing the intensity of the test utterances by the appropriate number of steps on a dB attenuator. Interfering speech and test utterances were mixed upon playback.

The onset to onset time of the test utterances on the stimulus tape was seven seconds, whereas the average duration of the test utterances was slightly over 1.5 s. Each utterance was preceded by a clear warning signal, consisting of a tone burst occurring 2 s before the onset of the utterance. In the actual experiment the stimulus sequence proper as described above was preceded by a block of 24 utterances in order for the subjects to get used to the monotonous artificial speech and the experimental situation. Ten employees of our institute served as subjects. They had some experience in listening experiments with resynthesized speech, but had not heard any of the test utterances before, and were unaware of the purpose of the experiment. Subjects listened individually, in a sound-insulated booth, to the stimulus tape, and were instructed to say aloud after each test utterance what they had heard. They were encouraged to repeat as much of the test utterance as possible where they had not heard all of it. Their voices were taped on one track of a magnetic tape, with the unmasked test utterances on the other track, for later analysis.

Results

In analyzing the results we calculated the percentages of errors against content words in each condition. Word substitutions in which only one phoneme was incorrectly reproduced were not included. This was done because it turned out that these errors had no relation with the experimental conditions and also occurred in a control test without interfering speech. The remaining errors were classified as substitutions and omissions of words.

The data show two major effects on the percentage of errors (substitutions plus omissions), one of the difference in pitch between target and interfering speech, and one of the difference in intensity. In a three-way analysis of variance, with subjects as the third variable, both effects were significant (Table I).

In Fig. 1 the percentages of errors, averaged over intensity levels and all ten subjects, are plotted as a function of Δ pitch, separately for (a) all errors combined, (b) substitutions only, and (c) omissions only. When the pitches of target and interfering speech are equal, the total percentage of errors is still only 60%. Apparently, even in the worst conditions, where perceptual fusion of target and interfering speech is maximized, intelligibility is not completely lost. The percentage of errors decreases with increasing difference in pitch, from about 60% for equal pitches to about 40% for a difference of three semitones, and probably would have decreased more with a further increase in pitch difference. When Δ pitch equals one octave, the percentage of errors is again relatively high, as expected on the basis of the inseparability of the harmonics of target and interfering speech, leading to perceptual fusion of coinciding voiced portions of the two messages.

At first sight it seems as if Δ pitch has a considerable effect on omissions, but has no effect on the percentage of substitutions. This, however, may be an illusion. Whenever for particular portions of the target speech perceptual fusion occurs with the interfering speech, the listener will most probably hear something unrecognizable and therefore may find it hard to come up with any response, correct or incorrect. We may interpret the decrease in percentage omissions with increasing difference in pitch as the effect of a decrease in the amount of perceptual fusion. Whenever the target speech does not fuse with the interfering speech, the recognition process is not inhibited and the listener can come up with a correct or incorrect response. The total number of cases where either correct or incorrect recognition can take place equals the total number of word presentations minus the number of omissions. Expressing the number of substitutions as percentages of the total number of word responses instead of the total number of word presentations indeed shows some effect of pitch difference: these percentages decrease from 43% for a Δ pitch of zero semitones to 35% for a Δ pitch of three semitones. Apparently, a difference in pitch between target and interfering speech affects the probabilities of both omissions and substitutions.

Figure 2 displays the percentages of errors, averaged over pitch conditions, as a function

Table I Table of results of a three-way analysis of variance applied to the data of Experiment I

	df_1	df_2	F	P
Pitch	5	423	3.18	<0.01
Intensity	3	423	12.6	<0.0001
Subjects	9	423	1.48	<0.15
Interaction pitch–intensity	15	423	1.05	n.s.

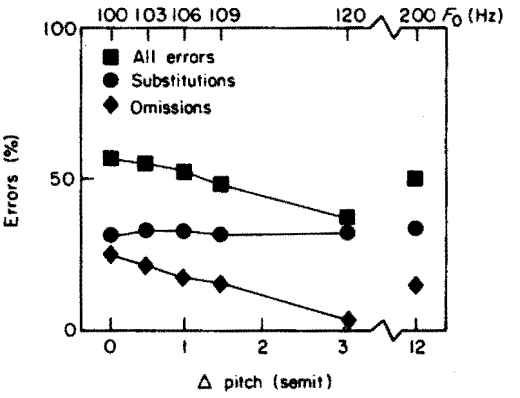

Figure 1 Per cent word errors as a function of the difference in pitch between monotonized target and monotonized interfering speech. The percentages are given separately for all errors combined, substitutions, and omissions.

of the intensity ratio between target and interfering speech, again separately for all errors combined, substitutions, and omissions. If, in this case too, we express the number of substitutions as percentages of the total number of word responses instead of as percentages of the total number of word presentations, we see an increase from 32% at 0 dB to 47% at − 15 dB.

The present results have shown that the intelligibility of speech in the presence of interfering speech from the same speaker is not only affected by the relative intensity between target and interfering speech, which seems more or less trivial, but can also be manipulated by varying the size of an artificial constant difference in pitch between the two speech streams. When the target and interfering speech have identical pitches throughout, the listeners are still able to recognize a sizeable amount of the words in the target speech, as it were through the holes in the interfering speech. When the pitches of the two speech streams are made different, the amount of recognized words increases, probably because the pitch difference helps the listener in perceptually separating the two voices. Of course, in a more realistic situation, the pitch contours of target and interfering speech will fluctuate

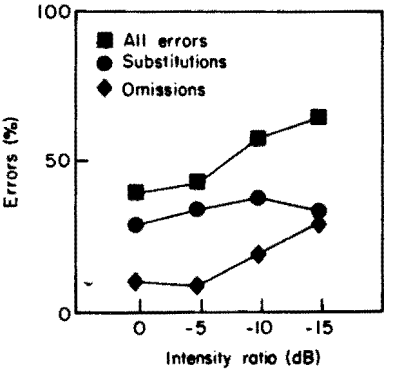

Figure 2 Per cent word errors as a function of intensity ratio between monotonized target and monotonized interfering speech. The percentages are given separately for all errors combined, substitutions, and omissions.

heavily, and if they are in the same pitch range, often cross each other. The effect of a difference in pitch range in such a situation is the main issue of the next experiment.

Experiment II

This second experiment was set up to examine the effect of the presence or absence of a difference in overall level of pitch between two speech messages, firstly for utterances spoken with normal intonation and secondly for monotonously spoken utterances.

Method

In this experiment the interfering speech was obtained in the same way as in Experiment I, except that the speech was not processed by a vocoder, and thus retained its normal intonation. The average pitch was about 110 Hz. The sentences used for the test utterances were of the same type as those in Experiment I, and again 96 of such sentences were used.

These were spoken by the same speaker as used for the interfering speech, in one of four different ways, 24 sentences for each way of speaking:

(1) normal intonation, low pitch;
(2) normal intonation, high pitch;
(3) monotonously spoken, low pitch;
(4) monotonously spoken, high pitch.

The first of these four conditions, the one with normal intonation and low pitch, resulted from this speaker's normal way of reading these sentences. The average F_0 level was about 110 Hz, and the pitch contour, showing a pitch accent on each of the four content words in each utterance, typically followed a rise–fall–rise–fall pattern, superimposed on a declination line. The second condition, normal intonation and high pitch, was first realized by a female speaker with an average F_0 of 220 Hz, and then the male speaker for our experiment was asked to imitate the female utterances as well as he could with the same pitch. Although subjectively, in the opinion of a panel of listeners, his imitations were very successful, *post hoc* measurements of the pitch contours showed that his pitch movements covered a wide range, occasionally reaching the normal male level in the valleys of the pitch contours. His average F_0 level was roughly 160 Hz. Otherwise the pitch contours were of the same type as those in the low pitch condition. The monotonously spoken utterances were obtained by asking the same speaker to read the sentences aloud in a monotone, tracking with the pitch of his voice a constant tone supplied to him via earphones. This tone was a sinusoid with a frequency of 110 Hz for the low pitch condition, and 220 Hz for the high pitch condition. The resulting utterances sounded monotonous, although objective measurements of F_0 still showed some fluctuations. The average F_0 was close to 110 Hz for the low pitch and 220 Hz for the high pitch condition. The speech tempo was slightly slower and the intensity differences between content words and function words were considerably smaller than in the utterances with normal intonation.

The 96 utterances were divided into six blocks of 16 utterances, each block containing four utterances from each of the four intonation conditions. Before the utterances were copied on one track of a two-track stimulus tape, which had the interfering speech on the other track, the overall intensity levels of all utterances were adjusted in such a way that the peak intensities were kept within a range of 2 dBA. The overall intensity level of the combined test utterances was adjusted relative to the interfering speech in such a way that the overall subjective loudness of test utterances and interfering speech was roughly the same. This was defined as the 0 dB intensity ratio between test utterances and interfering speech, used for the first block of 16 utterances. During the experiment the intensity ratios were

set before each successive block of 16 utterances by means of an attenuator. The ratios used were 0, −3, −6, −9, −12, and −15 dB. The interfering speech and the test utterances were mixed before they were fed to the subject's earphones.

Because we felt that there were rather few sentences per condition, and therefore the results might be obscured by accidental differences between sentences, we prepared a second stimulus tape following the same procedure except that we took care that all sentences used in the experiment were now spoken in another intonation condition. Thus the total amount of utterances was 192, divided over two stimulus tapes of 96 utterances. Each sentence occurred once on each tape, in different intonation conditions.

As in Experiment I, the onset to onset time of the test utterances on the tapes was 7 s, and the average duration 1.5 s. Each test utterance was preceded by a tone burst, occurring 2 s before the onset of the utterance, and functioning as a warning signal. The stimulus sequence proper was always preceded by a block of 32 utterances in the presence of interfering speech, in order for the subjects to get used to the experimental situation and their task. Twenty university students served as subjects, ten for each stimulus tape. They were paid for their co-operation. Subjects listened individually, to the tapes through earphones, in a sound-insulated booth, and were instructed to repeat each test utterance aloud. They were encouraged to repeat as much of the test utterance as possible in case they had not heard all of it. Their voices were taped on one track of a magnetic tape, with the uninterfered target test utterances on the other track for later analysis.

Results

As in Experiment I, we calculated the percentages of errors against content words per condition, not including word substitutions in which only one phoneme was incorrectly reproduced. A three-way analysis of variance, with intonation conditions, intensity ratio and subjects as variables, showed a significant effect of all three variables (Table II).

We will first concentrate on the results obtained with the sentences spoken with normal intonation. Figure 3 represents the percentages of errors (omissions and substitutions combined) as a function of the intensity of the test utterances relative to the interfering speech. The data are given separately for utterances having the same pitch range as the interfering speech, and those having a different (considerably higher) pitch range. Obviously, different pitch leads to a considerably better intelligibility than same pitch. For most of the intensity ratios there are roughly 20% fewer errors with different than with same pitch, or to put it differently, to obtain the same degree of intelligibility the utterances with different pitch can have a more than 6 dB lower intensity than the utterances with same pitch.

In Fig. 4 the percentages of omissions and substitutions are plotted as a function of the intensity ratio. The percentages of omissions are based on the total number of word presen-

Table II Table of results of a three-way analysis of variance applied to the data of Experiment II

	df_1	df_2	F	P
Intonation	3	799	12.2	<0.0001
Intensity	5	799	43	<0.0001
Subjects	16	799	3.48	<0.0001
Interaction intonation–intensity	15	799	0.57	n.s.

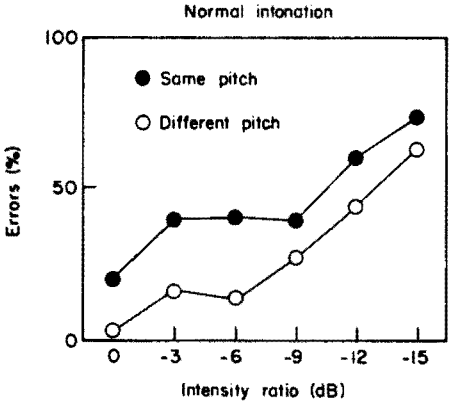

Figure 3 Percentage word errors as a function of intensity ratio between normally intonated natural target and interfering speech. The percentages are given separately for test utterances having the same pitch range as the interfering speech (SAME PITCH) and test utterances having a considerably higher pitch range (DIFFERENT PITCH).

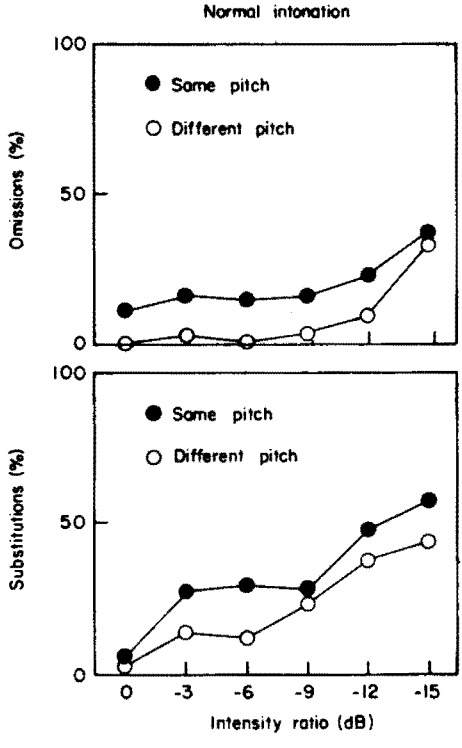

Figure 4 The same data as in Fig. 3, for omissions and substitutions separately. The percentages of substitutions are calculated after subtraction of the omissions from the total amount of errors.

tations, whereas the percentages of substitutions are based on the total number of word responses (i.e. number of word presentations minus number of omissions). This was done to avoid the obscuring effect of the number of omissions on the probability of substitutions. The percentages of both types of errors are systematically higher for the same pitch conditions, having the same overall pitch level as the interfering speech, than for the different pitch condition, having a higher overall pitch level than the interfering speech. This confirms, or at least supports, our idea that strongly overlapping pitch ranges of target and interfering speech are harmful to intelligibility, probably because whenever the two pitch contours come very close perceptual fusion may occur, and when the two pitches cross, the listener may get confused and track the wrong message.

In Fig. 5 the percentages of all errors combined, omissions plus substitutions, are plotted for the monotonously spoken utterances. These data are somewhat surprising. We had expected that the different pitch condition for monotonously spoken utterances would be extremely favourable to the intelligibility, because in this condition the pitches of target and interfering speech only very rarely come close to each other. The probabilities of perceptual fusion and crossing pitch contours are low. Comparing Figs 3 and 5, we see that in fact intelligibility in the different pitch condition for monotonously spoken utterances is rather low and comes close to the intelligibility of the same pitch condition for normally intonated utterances. Apparently, properties of the speech messages which are concomitant with a monotonous pronunciation, such as the absence of conspicuous pitch accents, and less prominent intensities of the content words relative to the function words, have unfavourably affected intelligibility, counteracting the favourable effect of separated pitch of target and interfering speech. This would explain the relatively high percentage of errors in the different pitch condition for monotonously spoken utterances. It does not, however, explain why there is so little difference between the different pitch and the same pitch conditions. We would still expect that the same pitch condition would show a systematically higher percentage of errors than the different pitch condition, due to the difference in probabilities of fusion and crossing pitches. Two possibilities suggest themselves here. One is that a monotonous pitch gives much less interference with normally intonated speech than a normal fluctuating pitch contour, for example because it is easier to track. This would explain why the same and different pitch conditions differ little in intelligibility. The other explanation would be that the combination of monotonous speaking and speaking at a pitch which was unnaturally high for our male speaker, made the speech in the monotonous high pitch condition (different) inherently less intelligible than the speech in the monotonous low pitch condition (same). In this way the favourable effect of different pitches would be counteracted by a relatively bad pronunciation. If the first explanation holds, we expect the two conditions to differ very little in the percentages of both omissions and substitutions. If the second explanation holds we expect that there will still be a considerably higher percentage of omissions in the same pitch condition, which in the percentage of all errors combined is counteracted by a relatively high percentage of substitutions stemming from the way of pronouncing. In Fig. 6 the percentages of omissions and substitutions are given separately as a function of intensity ratio. Again the percentages of substitutions are based on the total number of word responses and not on the total number of word presentations. There is clearly a systematic difference in the percentages of omissions between the two conditions, whereas there is no difference in the percentages of substitutions. Our interpretation is that, just as for normally intonated utterances, for monotonously spoken utterances too the probability of perceptual fusion and/or switching to the wrong message is decreased by separating the pitch ranges of target and interfering speech, but the favourable effect of

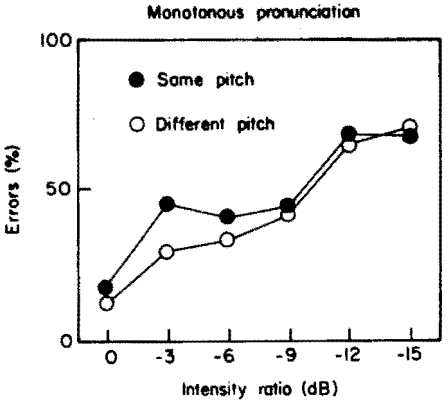

Figure 5 Percentage word errors as a function of intensity ratio between monotonously spoken natural target speech, and normally intonated natural interfering speech. The percentages are given separately for test utterances having the same pitch range as the interfering speech (SAME PITCH) and test utterances having a considerably higher pitch range (DIFFERENT PITCH).

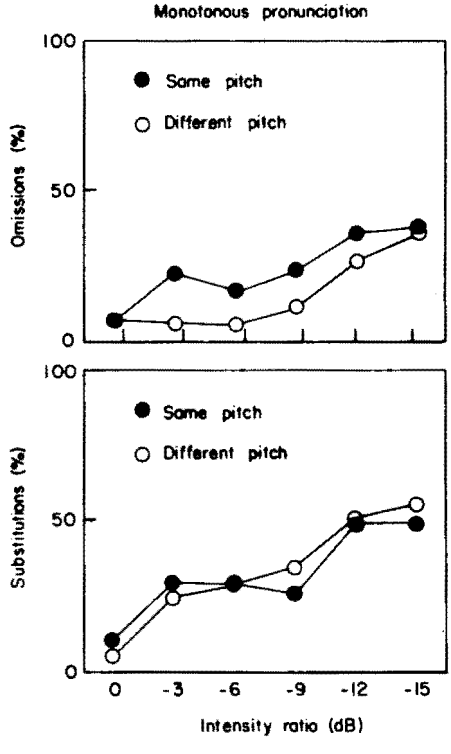

Figure 6 The same data as in Fig. 5, for omissions and substitutions separately. The percentages of substitutions are calculated after subtraction of the omissions from the total amount of errors.

this on overall intelligibility is counteracted by a non-intended difference in inherent intelligibility between the two conditions in this experiment.

Discussion

Perhaps the most impressive outcome of these experiments is the high performance of our listeners. Although they may have profited greatly from the fixed syntactic frame and the fixed number of syllables in the test utterances, the number of words they could choose from in each content word slot was vast, owing to the absence of semantic constraints. Yet, even at an intensity ratio between target and interfering speech of − 15 dB they could still reproduce correctly a sizeable proportion of the content words. Apparently, the listeners took advantage of the inevitable fluctuations in local intensity ratios and reconstructed much of the target utterances from what they heard through the silent or nearly silent gaps in the interfering speech. The main results of the present experiments show that the intelligibility of speech in the presence of interfering speech can be improved by introducing a difference in pitch between the competing speech messages. In a natural speech communication situation the disturbing effect of overlapping pitch ranges of competing messages is probably due to two separate effects, viz. the perceptual fusion of simultaneous portions of speech whenever the two pitches come close to each other, and the involuntary switching of attention to the interfering speech message, whenever the two pitch contours cross each other. In Experiment I we focussed on the first of these effects. This experiment demonstrated that, at least within a range of pitch differences from 0 to 3 semitones, intelligibility is a linear function of pitch separation. This favourable effect of pitch separation can best be understood from Goldstein's theory of pitch perception (Goldstein, 1973; Gerson & Goldstein, 1978). This theory assumes that the pitch of a sound is determined by a match between the harmonic structure of that sound and an internally generated harmonic template.

The "fundamental frequency" of this harmonic template is the perceived pitch. Two simultaneous sounds having the same pitch will be mapped onto the same harmonic template and thereby fuse into one sound. For two sounds having different pitches, two harmonic templates would have to be set up, which may help in sorting out two sets of harmonics from the stimulus sound, at least in the lower part of the spectrum. These two sets of harmonics, with their spectral envelopes, may then contribute to the perception of two distinct sounds. Thus, in listening to two simultaneous messages differing in pitch by a constant amount, the listener may be able to track the target message through those portions where both messages are voiced. This account of perceptual fusion of simultaneous voiced sounds is as yet highly speculative.

At present, the potential contribution of Goldstein's theory of pitch perception to understanding the perceptual separation of simultaneous speech sounds is further explored (Scheffers, 1979).

Whereas in Experiment I it was shown that a constant difference in pitch between competing messages has a favourable effect on the intelligibility of the target speech, in Experiment II we focused on a difference in overall pitch range, obtained by asking a speaker to speak either in his normal tone of voice or in a much higher tone of voice. Here we found a favourable effect of separating pitch ranges. In this case the effect can be due both to a decrease in the probability of perceptual fusion, and a decrease in the probability of switching attention to the wrong voice.

Our attempt to minimize the number of pitch crossings by asking the speaker to speak in a monotone and a high tone of voice, did not produce the high level of intelligibility we had expected. In the section on the results of Experiment II we have, on the basis of

the distribution of omitted and substituted words, argued that this unexpected result stems from pronunciation differences concomitant with normal and monotonous intonation. Further experimentation in this area should preferably use resynthesized speech with artificial pitch contours. In this way it would be possible to avoid undesired effects caused by differences in pronunciation, and to focus on the effect of pitch and pitch contours alone. Since the time we initiated the present experiments this has become much easier because of rapid developments in the quality and flexibility of computer-aided analysis and resynthesis of speech.

The use of resynthesized speech with artificial pitch contours could also help to clarify a further issue which we have left untouched in the present investigation, viz. the contribution of pitch contours, and in particular the fact that pitch contours are rule-governed and therefore to some extent predictable. It seems likely that the predictability of pitch contours helps a listener to continue tracking the target speech when the pitches of target and interfering speech cross each other. This is yet to be investigated.

Conclusions

The intelligibility of speech in the presence of other speech is better when the pitches or pitch ranges of the two competing speech messages are different than when they are the same. This effect can be related to the phenomenon of "perceptual fusion", occurring whenever two simultaneous sounds have identical pitches, and to "perceptual tracking": whenever the pitches of target and interfering speech cross each other, the listener runs the risk of inadvertently switching his attention from the target to the interfering speech.

This research was supported by a grant from the Netherlands Organization for the Advancement of Pure Research, as project no. 15-21-05. The authors are grateful to A. Cohen for the stimulating and inspiring way in which he helped to shape both the research and earlier drafts of this paper.

References

Broadbent, D. E. & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America*, 29, 708–710.

Brokx, J. P. L. (1979). *Waargenomen continuïteit in spraak: het belang van toonhoogte*. Unpublished doctoral thesis, Eindhoven University of Technology.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25, 975–979.

Cohen, A. & Hart, J. 't (1967). On the anatomy of intonation. *Lingua*, 19, 177–192.

Collier, R. (1972). *From pitch to intonation*. Unpublished doctoral thesis, University of Louvain.

Collier, R. & Hart, J. 't (1975). The role of intonation in speech perception. In: *Structure and Process in Speech Perception* (A. Cohen and S. G. Nooteboom, eds) Springer Verlag, Heidelberg, 107–121.

Darwin, C. J. (1975). On the dynamic use of prosody in speech perception. In *Structure and Process in Speech Perception* (A. Cohen and S. G. Nooteboom, eds) Heidelberg: Springer Verlag. pp. 178–193.

De Rooij, J. J. (1979). *Speech punctuation: an acoustic and perceptual study of some aspects of speech prosody in Dutch*. Unpublished Doctoral Thesis, University of Utrecht.

Egan, J. P., Carterette, E. C. & Thwing, E. J. (1954). Some factors affecting multi-channel listening. *Journal of the Acoustical Society of America*, 26, 774–782.

Gerson, A. & Goldstein, J. L. (1978). Evidence for a general template in central optimal processing for pitch of complex tones. *Journal of the Acoustical Society of America*, 63, 498–510.

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, 54, 1496–1516.

Hart, J. 't & Cohen, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics*, 1, 309–327.

Hart, J. 't & Collier, R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics*, 3, 235–255.

Myers, T. F., Zhukova, M. G., Chistovich, L. A. & Mushnikov, V. N. (1975). Auditory segmentation and the method of dichotic stimulation. In: *Auditory Analysis and Perception of Speech* (G. Fant and M. A. A. Tatham, eds) pp. 243–273.

Nakatani, L. H. & Dukes, K. D. (1973). A sensitive test of speech communication quality. *Journal of the Acoustical Society of America*, **53**, 1083–1092.

Nooteboom, S. G., Brokx, J. P. L. & De Rooij, J. J. (1978). Contributions of prosody to speech perception. In: *Studies in the Perception of Language* (W. J. M. Levelt and G. B. Flores d'Arcais, eds) pp. 75–109.

Scheffers, M. T. M. (1979). The role of pitch in perceptual separation of simultaneous vowels. *Institute for Perception Research, Annual Progress Report*, **14**, 51–54.

Stumpf, C. (1890). *Tonpsychologie*. Leipzig: S. Hirzel-Verlag. (Re-issued by Hilversum-Amsterdam: Knuf-Bonset, 1965).

Treisman, A. M. (1964). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, **12**, 242–248.

Van Katwijk, A. F. V. (1974). *Accentuation in Dutch: an Experimental Linguistic Study*. Assen: Van Gorcum.