

# From prominent syllables to a skeleton of meaning: a model of prosodically guided speech recognition

Robert Bannert

## ABSTRACT

In this paper an attempt is made to contribute to a better understanding of the processes involved in speech recognition. The results of some experiments constitute the basis for an outline of a model of speech recognition where prosody, especially the accentuated syllable, plays a guiding role.

In the experiments, Swedish utterances were used that were spoken with a foreign accent and that, above all, deviated tonally and rhythmically in a clear way. By introducing appropriate corrections step-by-step in a controlled manner (certain selected prosodic features in the speech signal representing the utterances were altered by LPC-synthesis), it was possible to observe the effect of these features on perception. It turned out that the word accent plays a predominant role. Accentuated syllables act as islands of clarity and stability in the speech chain due to their joint marking by a tonal change, clear spectrum, duration, and intensity. Therefore the listener is able to process these prominent and conspicuous syllables quickly and accurately. Other linguistic features (spectral, morphological, syntactic, and semantic) which are contained in the phonetically blurred and incomplete parts of the signal are subordinated to the linguistic structure of the identified accent pattern.

The model of prosodically guided speech recognition presented here shows some specific features. For instance, the processing units are not considered to be words in the orthographical sense (strings of letters separated by spaces on paper) but rather the accentuated syllables as anchors or fixation points surrounded by unstressed syllables. The speech recognition process does not lead to the identified meaning of a given utterance by matching a word completely analysed and with exactly defined boundaries by the acoustic-phonetic analysis with a corresponding template stored in the lexicon. The identification of meaning is instead achieved by a continuous interactive searching which is performed as an ongoing interplay between different levels of the speech recognition process. During this interplay, the current structure may be altered at any time as a consequence of several factors: new acoustic information which is continuously extracted from the incoming speech signal, linguistic constraints applicable to the intermediary structures, and the general and pragmatic knowledge of the listener.

## INTRODUCTION

Perception plays a very important part in speech communication. Perception in speech can be defined as the listener's active processing of the speech signal in order to reconstruct the message intended by the speaker.

There exist a number of models of word and speech recognition. They represent different approaches and views and partly are designed in basically different ways. Often word recognition is synonymous with pattern recognition implying that the structure arrived at by the acoustic-phonetic analysis of the incoming signal is compared and matched with templates that have been stored previously. As the next step, lexical access is executed where the semantic representation, i.e. the meaning is identified. Pisoni (1984) points out particularly that a clear distinction has to be made between word recognition and lexical access. Pisoni also presents a good overview and a review of seven recent models of word recognition, among which the Phonetic Refinement Theory developed by him and his collaborators (Pisoni et al. 1985) is to be found.

The Phonetic Refinement Theory and the model of Shipman and Zue (1982) represent a real development of earlier models because they take into due consideration phonetic features and interrelationships. However, it seems to me that these models can be further developed and supplemented, especially where the role of prosody, i.e. the tonal and rhythmic aspects of speech, in word and speech recognition is concerned.

Every linguistic unit, like syllable, stress group, phrase, sentence, and text, has a specific structure, the knowledge of which is of central significance for speech recognition. The competence of the speaker/listener also contains, among other things, the knowledge of the phonotactic structure of syllables and words, their morphological structure (root, affixes), their prosodic structure, and the number of reductions and assimilations. The prosodic features are very often strongly interrelated with other phonological and morphological features, for instance phonotactic, morpho-phonological, and syntactic ones.

Models of speech perception have to cope with the fact that the speech signal is not always distinct and complete. Instead, most often the acoustic signal arriving at the listener's ear contains distortions of different kinds. These

deviations appear as the consequences of at least three dimensions of indistinctness, namely of speech tempo (slow - fast), of articulation (distinct - lax), and of the linguistic distance between a norm or standard and the actual form (small - large) which contains regional, social, and individual features and foreign accent as well. These deviancies include reduced or missing spectra compared to the intended form both of which can originate from lax or fast speech or from external distortions. Opposite to incompleteness, the signal may contain a larger number of segments, e.g. produced by vowel epenthesis. Some spectra may be analysed only to a certain extent, e.g. an [m] only as nasal, an [ø] only as palatal, prosodic features may be wrong (short instead of long vowel), the accent may be placed on the incorrect syllable (telephóne), etc. Therefore it has to be assumed that the result of the acoustic-phonetic analysis not always amounts to a complete and unambiguous phonological form which will lead directly to the lexical element which, eventually, will be identified correctly. On the contrary, the phonological representation as the result of the working of the bottom-up processes has to be thought of as incomplete and deviant compared to the meaning intended by the speaker.

An adequate model of speech perception should be able to handle a rather wide variation in the speech signal. Listeners do communicate with each other in spite of individual, social, geographical, and other differences in their pronunciation. A clear instance of large acoustic deviances is to be found in foreign accent. Thus phonetic variation is a rather complex phenomenon, resulting from a given language being spoken not only as the first language but also as the second language by people with different first languages. A simplified multi-dimensional model of phonetic variation is presented in Bannert (1982).

It is the aim of this investigation to present an outline of a model of speech recognition in which prosody plays a significant and leading role (\*). This model is compatible, to certain parts, with other models of word or speech recognition, especially with the Phonetic Refinement Theory presented by Pisoni et al. (1984). However, it adds some new aspects focussing on the incompleteness and fuzziness of the results of the acoustic-phonetic analysis. It does not work with the processing unit of the word characterized by clearly defined boundaries. The phonological form of the word is being built and assembled starting from phonological fragments and applying, among other things, the morpho-phonological knowledge of the listener.

## METHOD AND MODEL

As the basis for testing intelligibility of Swedish spoken with a foreign accent, the so-called correction method is used (Bannert 1978). The starting point is an utterance spoken by a non-Swedish speaker. It contains certain features which are analysed and well-defined. These deviations are corrected in the acoustic signal step-by-step by means of LPC-synthesis. The tonal feature of accent corresponds to certain parts of the Fo-contour of the utterance, the temporal features of quantity and phrase rhythm are manifested in the durations of the segments and their relationships to other segments and groups of segments (syllables). The Lund model for Swedish intonation (Bruce 1977, Gårding and Bruce 1981) served as the model for the tonal corrections. The temporal corrections had to be carried out according to estimated values which were then checked auditorily. We are still lacking a comprehensive model for Swedish speech rhythm.

Intelligibility of foreign accent is investigated against the background of a kind of Active Direct Access Model for word recognition similar to the model developed by Marslen-Wilson and Welsh (1978). Recognizing a word is considered a time-dependent active process where acoustic (bottom-up) and linguistic, pragmatic, and general information (top-down) conspire.

Intelligibility is seen as an aspect of the bottom-up processes. Intelligibility is high if the acoustic, auditory, and phonetic analysis of the speech signal results in a possible phonological structure processed in the short-term memory that easily and quickly can find its way to the phonological representation of a word stored in the long-term memory. In the opposite case, intelligibility is low if the speech signal is analysed in such a way that no corresponding lexical element can be discovered. Thus intelligibility facilitates decoding by decreasing the demands on the top-down component and makes comprehension faster, easier and better.

If the model of Active Direct Access is expanded to foreign-accent speech, one prediction, then, will be that a longer stretch of acoustic information - a larger chunk of the speech signal - is needed before a word spoken with a foreign accent can be recognized. Thus, due to the acoustic deviations, it should also take more time to process foreign accent. Foreign accent puts a lot of strain on the short-term

memory and a heavier demand on the top-down processes.

## TEST PARADIGM

Testing intelligibility is not an easy task. After considering different problems, their approaches and possible solutions, a test paradigm was constructed which is shown in Fig. 1. Samples of foreign accent that are clearly deviating in the prosodic features to be investigated constitute the starting point. Using LPC-speech synthesis, each utterance is altered in such a way that corrections corresponding to the deviating prosodic features are introduced into the speech signal <1>. In this way, families of utterances are created that consist of several members, namely the original utterance and several versions that differ from the foreign accent original by a certain correction or improvement. Each family of utterances is extended by adding an idiomatic version spoken by a male Stockholm speaker. Thus a dimension of variation within each family is established where the foreign accent original and the Swedish version mark the end points and the corrected versions are assumed to lie in between.

All these utterances are then distorted in different ways, using noise and increased speech tempo <2>. This test design makes it more difficult for the listener to understand speech, and, at the same time, it is easier to discern the effects of the various corrections. In both tests, the signal-to-noise ratio in the noise test and the speech tempo in the second test were chosen based on preliminary tests using naive listeners. The listeners understood the utterances with difficulty.

The Swedish listeners who were not accustomed to foreign accent participated individually in the listening tests. They heard the test utterances via loudspeakers in the perception laboratory and repeated in their own Swedish without hesitation what they could understand of the utterances played to them <3>. The listeners were urged to respond, even if they did not understand the whole utterance and to guess freely in case of uncertainty. The test utterances and each listener's responses were recorded on different channels of a REVOX tape recorder. The responses were analysed, evaluated, and compared to the intended meaning. Response time in the noise test was measured. Transmitted prosodic information analysed in the oral responses and response time are the

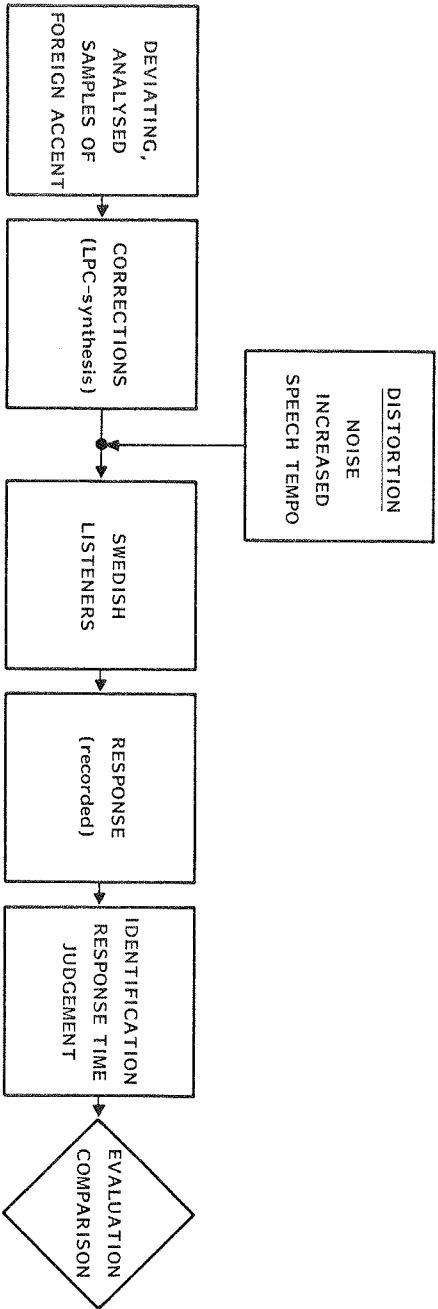


Fig. 1 The test paradigm.

basis for a ranking of the phonological features.







## MATERIAL AND CORRECTIONS

The test material consisted of 9 relatively short utterances <4> the length of which ranged from a single compound word of 3 syllables (utterance 6) to a complete sentence consisting of subject, adverb, and adverbial phrase (utterance 4). The test utterances were rendered by three male speakers with French, Greek, and Persian (Farsi) as their first language (L1). These utterances and their corrections are shown in Fig. 2. The features of quantity and phrase rhythm are corrected by changing segment durations, the feature of accent <5> is corrected by inserting a tonal peak in the accentuated syllable and, at the same time, deleting the original peak in the incorrect syllable <6>. An illustration of the temporal and tonal corrections in utterance 8 is shown in Fig. 3.

The kind of stimuli and their number varied from test to test. The test material common to all 3 tests consisted of the 9 original utterances spoken with foreign accent and three corrections each. They were interspersed with twelve different utterances spoken with foreign accent which served as distracters and, at the same time, as calibrators for the reliability of the listeners' responses. The intelligibility tests also contained a version of each of the original utterances of foreign accent spoken by a male Stockholm speaker. Furthermore, the intelligibility test presented with increased speech tempo and the acceptability test also contained a version of each of the 9 original foreign accent utterances representing deteriorated Swedish. These deteriorated Swedish utterances were produced by introducing into each of the original Swedish utterances all the prosodic deviations of its original utterances <7>. All the stimuli were re-synthesized. They were free from distortions such as clicks or buzzes and sounded quite natural.

## LISTENING TESTS

Each listening test concerning intelligibility consisted of three parts. First, the utterances of the four speakers were presented in the following order: French, Greek, Persian, and

UTTERANCE No	LI	UTTERANCE	CORRECTIONS
1	FRENCH	EN KAFFEBRICKA	  
2	"	SOM EN MYCKET LITEN POTATIS	
3	"	LITE RÖDA TYGBITAR	
4	GREEK	Å SOLEN LYSER BLEKT I SÖDER	  
5	"	BÅDA ÄR DYRA	
6	PERSIAN	MARKATTA	
7	"	DET ÄR EN MÅNDAGMORGON	
8	"	I SAMHÄLLET	




 PITCH ACCENT  
 PHRASE RHYTHM  
 QUANTITY

Fig. 2 The eight utterances and their corrections.

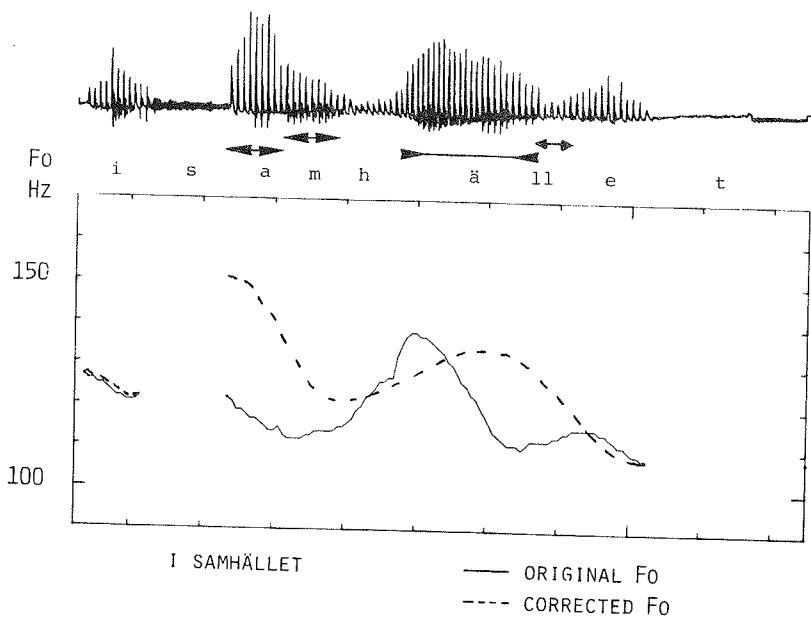


Fig. 3 An illustration of the temporal and tonal corrections.



Swedish. The speakers read aloud a text, approximately 45 seconds each. Thus the listeners were given the opportunity to adapt to the four speakers' voices and pronunciation. This speaker adaptation is also necessary as a precondition in order to be able to comprehend speech in a normal way. In the acceptability test, the Swedish speaker was excluded, as were the original Swedish utterances. Second, eight utterances (six in the acceptability test) followed, two for each speaker, presented in the same way as the test proper. The utterances of the second part were intended to get the listeners accustomed to the presentation of the utterances under noise, with increased speech tempo for the intelligibility tests or in the normal way for the acceptability test and to practice responding to the stimuli. These practice utterances did not appear in the test. Third, the noise test proper contained 21 utterances, namely nine test utterances (1 version of each original utterance) and 12 distracters. As each listener responded to each utterance only once, the 45 stimuli [(original foreign accent utterance + 3 corrections + original Swedish version) x 9 utterances] of the whole test were divided into 5 series and administered separately to 5 different listener groups. The test with increased speech tempo contained 54 test stimuli (45 stimuli + 9 deteriorated Swedish utterances). The test was divided into 6 series and given to 6 new listener groups.

In the two intelligibility tests, each listener heard and responded to each test utterance only once. In the acceptability test three marks (1 for a low degree of foreign accent, 2 for a high degree of foreign accent, and x for a degree of foreign accent in between) were given to each stimulus by each of the 20 speakers. The total duration of the listening test under the conditions noise and increased speech tempo, respectively, was about seven minutes. The acceptability test took about 20 minutes. Fifty South Swedish listeners (university students) took the noise test individually. Thus each version was given 10 responses by the whole group. The test with increased speech tempo was taken by 30 listeners under the same conditions. Thus, in this test, each version was given 5 responses by the whole group.

The responses of the two intelligibility tests were analysed and evaluated according to a scoring system that primarily counted the prosodic information contained in the listeners' responses, such as number of accents, number of syllables (vowels), stress pattern, syllable quantity (long/short vowel and consonant), etc. The response time was defined as the time lag between the end of the test utterance and the beginning of the listener's response measured on

duplexoscillogrammes. The marks of the acceptability test were counted and are given as group scores.

## RESULTS

The results of the three experiments are reported as follows: The effects of various prosodic features along the dimension intelligibility are shown under the two conditions, namely original speech tempo and noise in Experiment 1 (Fig. 4) and increased speech tempo in Experiment 2 (Fig. 5). In parallel and supporting this aspect, the reaction times of Experiment 1 illustrate rather the psycholinguistic dimension of speech processing (Fig. 6). Some selected typical examples of listener responses that did not correspond to the intended utterances and which may provide some revealing information about the processes involved in speech recognition are presented and analysed. The assessment of the stimuli according to their acceptability in Experiment 3 provides some useful illustration of the psychological and subjective aspects of speech recognition (Fig. 7). Finally the results of each experiment are compared with one another: Intelligibility and Reaction time in Experiment 1, both with Intelligibility in Experiment 2, and the three of them with Acceptability in Experiment 3.

### Intelligibility

In Figures 4 and 5, the distribution of the stimuli under the condition Noise (original speech tempo) and Increased speech tempo, respectively, are given as percentage along the dimension intelligibility which is defined in terms of transmitted prosodic information and represented as a straight line. Each of the 8 families of utterances <4> is shown individually.

Fig. 4 shows that the utterances with the original and uncorrected foreign accent often have only a low intelligibility which expectedly is in opposition to the Swedish corresponding utterances that are understood very well and without difficulty. However, in both cases there are several exceptions. Utterance 4 is especially conspicuous showing a relatively high degree of intelligibility in spite of its foreign accent. The corresponding Swedish utterance, on the contrary, shows a low degree of intelligibility. In

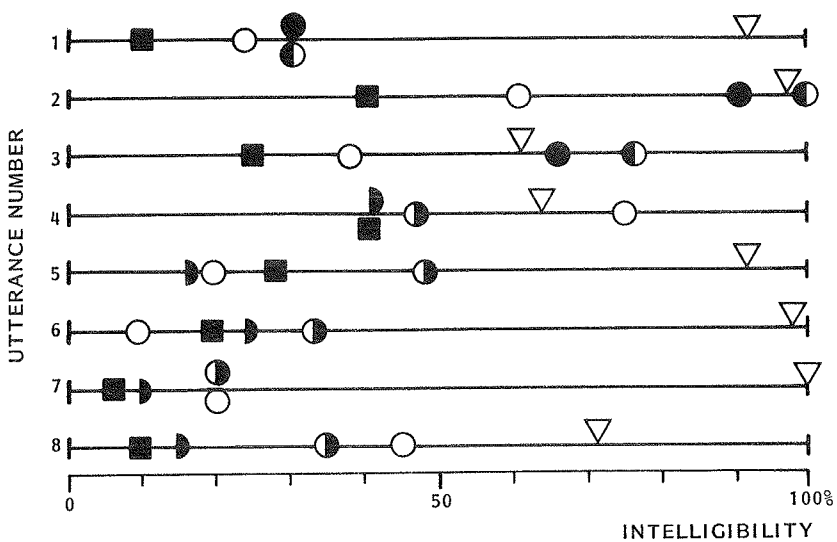


Fig. 4 The effect of the corrections on intelligibility expressed as the prosodic information contained in the listeners' responses (noise condition).

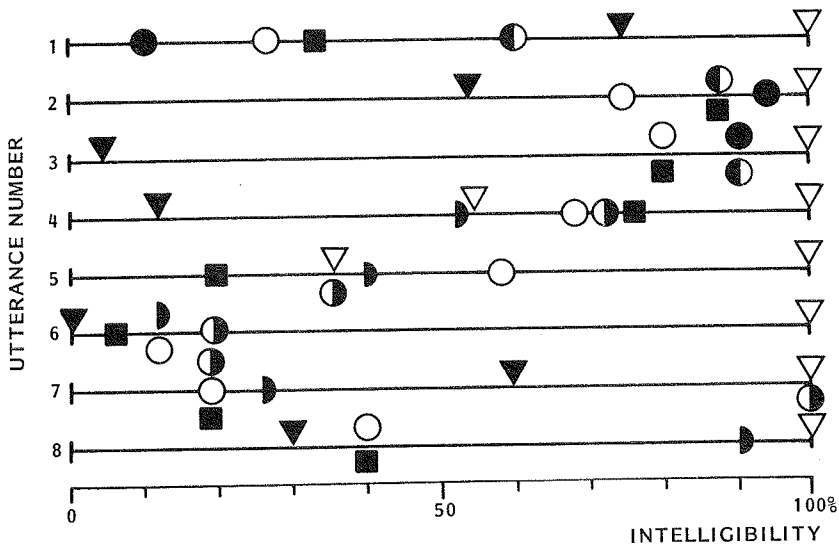


Fig. 5 The effect of the corrections on intelligibility expressed as the prosodic information contained in the listeners' responses (increased speech tempo).

most utterances, the corrections bring about an increase of the degree of intelligibility, utterances nos. 4, 5, and 6 being the exceptions. In some cases (utterances nos. 2, 3, 4), in fact, a high degree of intelligibility is reached as a consequence of the corrections. Distributed across the whole material, the corrections of word accent and (phrase-)rhythm produced the best results (exceptions here are utterances nos. 4 and 8).

Fig. 5 also shows the distribution of the stimuli along the dimension Intelligibility, but here the deteriorated Swedish versions are added. Compared to Fig. 4, this result, by and large, is quite similar. The utterances with the foreign accent show a relatively low degree of intelligibility, their original Swedish counterparts, on the other hand, a very high degree. Even under this condition, utterance no. 4 is the exception. The original Swedish utterances suffer from a dramatic decrease in the degree of intelligibility when the deterioration of the prosodic features are introduced into the signal. This holds especially for utterances nos. 3, 4, and 6. Under the condition of Increased speech tempo, too, the corrections increase intelligibility. And here, too, the combined correction of word accent and (phrase-)rhythm lead to the best results in most cases (clearly in utterances nos. 2, 3, 4, 6, 8).

A comparison of Figures 4 and 5 reveal minor differences. For instance, the corrections of utterances nos. 2, 3, and 8 in Fig. 5 (Experiment 2) produce higher values. These differences might be attributed above all to the different conditions in the two experiments. The behaviour of utterance no. 4 which clearly deviates from the other utterances, even with respect to reaction time (see the following section), might be attributed to its syntactic complexity and its length.

### Reaction times

Fig. 6 shows the distribution of the stimuli of Experiment 1 (original speech tempo, noise) according to reaction time of the group scores. Compared to the results concerning intelligibility, by and large, similar relationships between the different versions of an utterance are to be found. The original Swedish utterances almost always show the shortest reaction times, as could be expected. But even in this case, utterance no. 4 comes out as the real exception. In only a

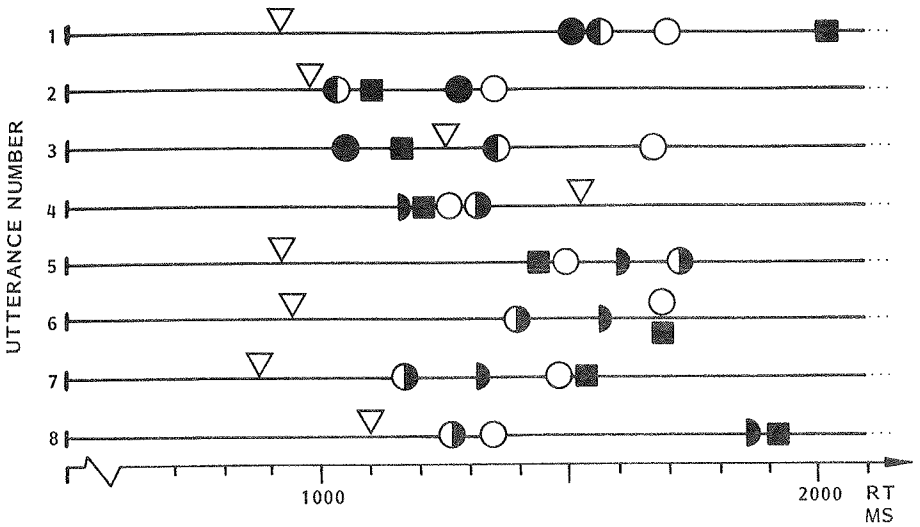


Fig. 6 Response times (noise condition).

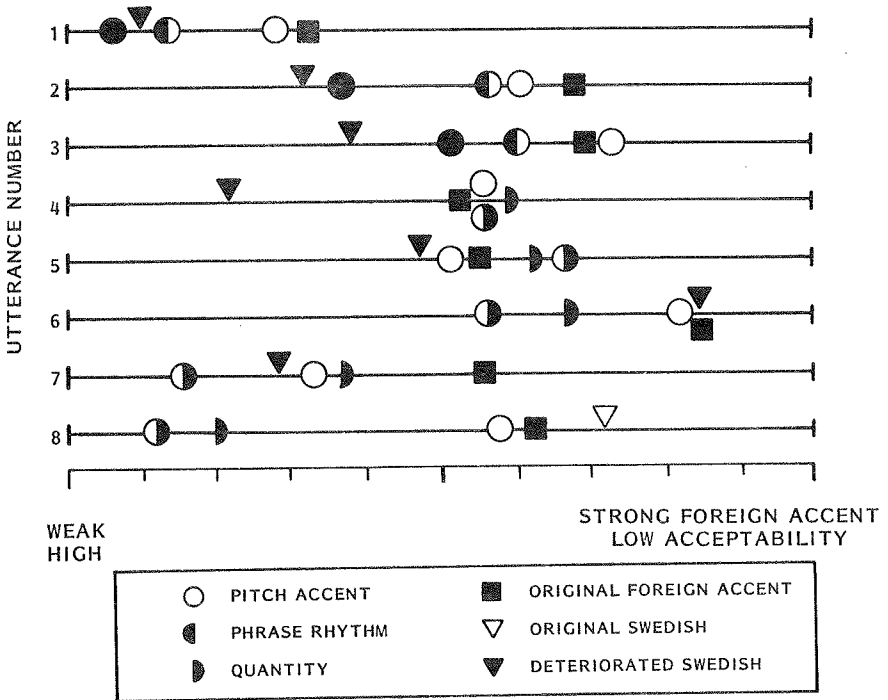


Fig. 7 Acceptability of the original foreign accent utterances, their corrections, and the deteriorated Swedish utterances.

few cases do the uncorrected foreign accent utterances have the largest reaction times. The shortest reaction times among the corrected utterances are for those stimuli in five cases where word accent and (phrase-)rhythm were corrected in combination (utterances nos. 1, 3, 6, 7, 8).

A comparison of these results concerning reaction times in relation to the corrected versions reveals that the combined correction of word accent and (phrase-)rhythm are characterized by the shortest reaction times which may be interpreted as an indication of easy and fast processing and, at the same time, bring about the highest degree of intelligibility.

### **Acceptability**

Fig. 7 shows the distribution of the stimuli assessed by the listeners according to the degree of foreign accent which may serve as a direct measure of acceptability. The test material here is arranged along the dimension of high vs. low acceptability. In most cases, the utterances with the original foreign accent are assessed with the lowest degree of acceptability. Only utterances nos. 4 and 5 represent clear exceptions. Those corrections where the features of word accent and (phrase-)rhythm were manipulated in combination, with only one exception, namely utterance no. 5, show the highest degree of acceptability.

The deteriorated versions of the original Swedish utterances show a very low degree of acceptability in six cases (utterances nos. 1, 2, 3, 4, 5, 7), and in four cases (utterances nos. 2, 3, 4, 5), in fact, the lowest degree of all versions.

Comparing the results concerning acceptability with those concerning intelligibility under different conditions, reveals a clear and parallel behaviour of certain stimuli. Among the corrected features, the combination of word accent (accent pattern) and (phrase-)rhythm stands out as the most efficient one. These stimuli obtain the highest degree of intelligibility under different conditions, need the shortest reaction times, and are accepted most readily.

## Response patterning

Analysing and evaluating listener responses with respect to the altered features of the stimulus may provide some information about the processing of the incoming signal. Those cases where the listener did not respond at all or responded with the intended utterance <8> are not very interesting.

Some typical examples of listener responses showing the effects of only correcting word accent on the results of the speech recognition processes, i.e. putting the tonal change onto the correct syllable, are given in Table 1.

In general, the accentuated syllables in the stimulus, no matter where, come through very well. The responses to the original utterances where word accent falls on the wrong syllable correspond exactly to this accent pattern. By correcting the word accent only, i.e. by shifting the tonal change onto the right syllable, the responses change in such a way as to correspond to the new and correct accent pattern. In many cases it can be observed that the accent pattern of the incoming signal determines the accent pattern of the response which will be identical, although the other linguistic structures and features of the response differ with respect to the stimulus. It seems as if spectral, morphological, syntactic, semantic, and pragmatic elements are fitted into the framework laid out by the accent pattern. Thus it appears rather clearly that the word accent syllable serves as the first and decisive sign post or guide in the processing of the acoustic information at non-peripheral levels. Therefore word-accentuated syllables, as a consequence of their prominent marking by combining tonal, rhythmic, spectral, and dynamic features in them, play a predominant part in speech recognition. Other linguistic aspects of the possible linguistic structure, drawing upon all kinds of information available, seem readily to be subordinated to the gross structure defined by the accent pattern.

## DISCUSSION

First the effects of the corrected prosodic features on speech recognition are commented upon. Second an outline of a model of prosodically guided speech recognition will be

Table 1. Some typical listener responses

SWEDISH	FOREIGN ACCENT (ORIGINAL)	CORRECTED WORD ACCENT
en 'kaffe,bricka	(en kaffe'bri'cka:) ja e inte 'klar en liten 'flicka	en 'vacker 'flicka 'kaffet e 'klart
'båda är 'dyra	(båda är 'dyra) va de 'blir på dig 'båda fotogra'fi	'båda,dera 'både ...
'mar,katta	(mar'ka:ta) man 'pratar dom e 'korta ma'kaber	'prata 'marknad
det är en 'måndag,morgon	(det är en mån'da:gmorgon) det är en nou'gatmålning, det är en han'garmålning	det är en 'vårmorgon, det är en 'söndagmorgon, det är en 'kundradio
i 'sam,hället	(i sam'hä:let) i sin 'helhet utan 'teve	i 'sandträdet i 'samlingen i 'handlingen

The accentuated syllable is marked by ' preceding it. A long vowel is specified in the very broad transcription of the original stimulus in parentheses for reasons of clarity.



presented and, third, speech recognition is discussed under the aspect of the difficulties related to foreign accent.

### **The effect of the corrected prosodic features on speech recognition**

It was expected that the manipulations in the speech signal which were made in a controlled and step-wise way should make it possible to make clear and definite statements about the effect of the corrected prosodic features. However, the results clearly show that there is not always a simple and direct relationship between the correction of a given prosodic feature and the listeners' reaction to it. Thus it happens several times that the corrected version shows a lower degree of intelligibility than the utterance with the original foreign accent (for instance Fig. 4, utterance no. 5; Fig. 5, utterances nos. 1, 2, and 4). Corresponding statements can be made with respect to reaction times and acceptability. In the opposite case, the Swedish original utterances, too, get low scores rather often and, in fact, they score worse than some corrected versions (for instance, Fig. 4, utterances nos. 3 and 4; Fig. 5, utterance no. 4; Fig. 6, utterance no. 4).

This unexpected behaviour may have several explanations. No doubt, however, the reason can hardly be found in the fact that the score of each version in the dimensions Intelligibility, Reaction time, and Acceptability, respectively, was given by a different group of listeners. In order to eliminate this supposed factor, the number of the listeners has to be increased considerably. But, as far as the assessment of the speech signal by the listeners is concerned, it is quite clear that phonetic and phonological deviating features are analysed and evaluated rather differently. In my experience, this individual reaction pattern on behalf of the listeners always becomes obvious, even when trained and experienced teachers of Swedish as a second language are exposed to foreign accent.

However, a more plausible explanation for this divergent behaviour seems to be found on purely phonetic and phonological grounds. The various manipulations, e.g. only vowel or consonant duration, only word accent, may have a somewhat negative effect on the processing of the speech signal by the listener. This is because some manipulations may interfere with the various processes involved in speech recognition or even impair and, at worst, block them. We must

not forget that even the corrected signal is clearly heard as foreign accent as it still contains certain prosodic and all of the segmental deviances. By correcting only one feature at a time, new constellations of foreign accent may be created which, against the background of the interplay between prosodic features, segmental features, morphological and syntactic features, may exert an impairing influence on the processing of the speech signal.

In conclusion, then, a general statement can be made: compared to the original Swedish utterances and also to the deteriorated Swedish utterances, most of the prosodic features affect speech recognition in a positive way by increasing the degree of intelligibility, by decreasing the reaction time, and by being accepted to a higher degree compared to their deviant counterparts. The largest positive effect is exerted by the combination of word accent and rhythm.

With respect to the significance of prosody in speech recognition, another general statement can be made: The accent pattern, rhythmic structure, and overall intonation contour facilitate purposefully the successful processing of the speech signal. These features give a macro-structure to the speech chain by dividing the spectral events or the stream of sounds into useful units larger than sounds and syllables, namely accent groups, prosodic phrases or intonation units (cf. Nespor and Vogel 1983) <9>.

#### **Outline of the model**

On the basis of the results of my investigations and the models of word recognition mentioned in the introduction, a prosodically guided model of speech recognition has been developed. Prosodic features play the decisive part for the searching of lexical elements. The model, outlined in the following in a simplified way, is shown in Fig. 8 <10>. It describes an interactive process on several levels where information and knowledge of various kinds affect the recognition process from the speech signal to the identified meaning.

Starting with the input, the acoustic-phonetic basic information of (one part of) the utterance is extracted by the peripheral auditive-acoustic analysis. This first automatic analysis proceeds from left-to-right, i.e. the incoming speech signal is processed continuously along the

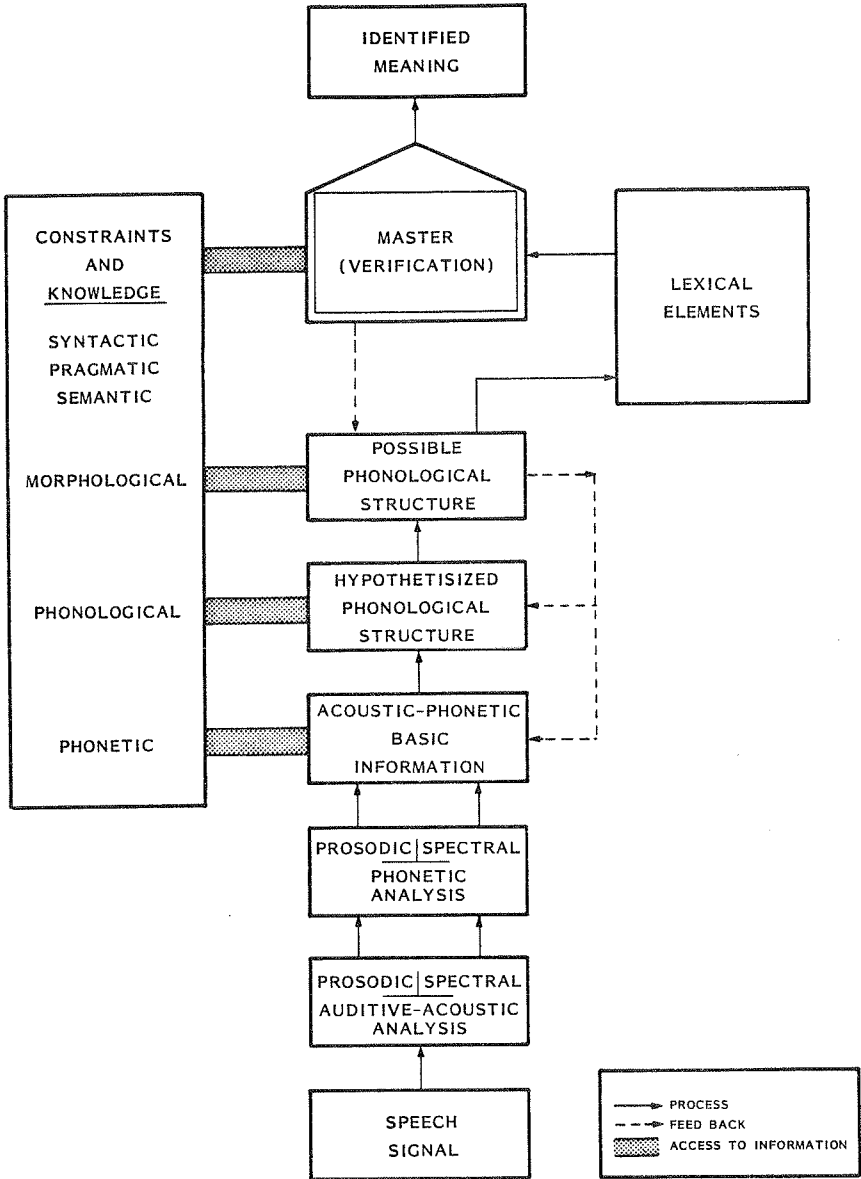


Fig. 8. A model of prosodically guided speech recognition.

time axis.

The acoustic analysis is done in two different channels, namely the prosodic and the spectral one (cf. Svensson 1974, House 1985). While in the prosodic channel the tonal and temporal features of the chunks of the processing units are established, the spectral channel provides the information about the qualitative features (formant frequencies, band widths, etc.) of the segments <11>.

Quite often the auditive-acoustic analysis cannot always result in a complete phonetic basic structure. The speech signal may be blended with acoustic distortions from outside, the signal-to-noise ratio may be too small or the speech signal might contain in some form components that are reduced, missing or, with respect to the form expected by the listener, deviant in some other way.

The auditive-acoustic analysis is followed by the phonetic analysis which combines and integrates the auditive-acoustic parameters into chunks of approximately the size of a syllable and which labels it phonetically. The phonetic labelling, most often, cannot be performed in a refined way (cf. Pisoni et al. 1984). The phonetic interpretation provides the basis for the acoustic-phonetic basic information about the chunk of the speech signal to be processed.

The acoustic-phonetic basic information is structured according to prosodic and spectral features. The prosodic features provide the position of the accentuated syllable or syllables in the chunk or chunks; the spectral features contain information about the spectral gestures of the segments. Taken together they provide information about the number of syllables in the chunks. There is, however, a clear difference between the two dimensions: while the accentuated syllable always appears correct in the basic structure, the spectral component often remains classified only in a gross manner.

This fact has certain consequences for the emergence of the hypothesized phonological basic structure on the following level: The spectral elements in the acoustic-phonetic basic information are subordinated to the prosodic structure of the accent groups where accent group means the accentuated syllable surrounded by the unstressed syllables. This subordination is brought about by the top-down constraints and the general knowledge of the listener which operate in generating the hypothesized phonological structure. These

constraints are phonetic, phonological, morphological, syntactic, and semantic.

The hypothesized phonological structure is not generated only once and for ever but, instead, can be altered in a short period of time as a consequence of not only new acoustic-phonetic information but also of new top-down information which is flowing forth and thus becomes available all the time. The definite hypothesized phonological structure of accent groups generates the possible phonological structure of (chunks of) utterances which are stored in the Short-Term Memory (STM) as well. Now the search in the lexicon in the Long-Term Memory (LTM) for lexical elements which correspond phonologically to the equivalent elements stored in the lexicon will start. The semantic elements of the lexicon are arranged in a multi-dimensional fashion according to various phonological features and structural characteristics. These possible phonological structures provided by the analysis of the speech signal and the working of linguistic constraints, it must be assumed, normally do not look like orthographic words with clearly defined boundaries, which correspond exactly to a stored counterpart. They are not searched for like a numbered book in a bookshelf and found immediately by its distinctive digit. Approaching the lexical elements would rather amount to a search consisting of a large array of activities utilizing different features simultaneously. The possible phonological structure which emerged from the fragments of the acoustic-phonetic basic information contains the accentuated syllable as its most important search criterion which is stuffed with the most distinct acoustic and structural information. Therefore it can be assumed that the search starts out for phonological representations of lexical elements showing the identical accent pattern and most of the spectral features of the accentuated syllable. Of course, all the information concerning the surrounding syllables is used as a supporting criterion as well. In general, it has to be assumed that speech recognition is characterized by an interplay of activities where all information available is processed simultaneously and optimally. This kind of search assumes explicitly that the boundaries in the possible phonological structure need not be defined exactly and in advance. The first aim of the search for lexical elements seems to be to find the syllables with the most distinct marking which, in turn, are identical with the basic meaning of the root or stem of a word, i.e. to find the skeleton or the corner stones of meaning.

As is generally known, languages use different principles for

accent distribution in their information structure. In accent languages like, for instance, Swedish, English, and German, word accent, in principle, exactly functions for signalling the word stem as the kernel of the meaning of a word. This is true both of morphologically simple and complex words. But also in languages with different principles for accent distribution, like for instance Finnish and Czech with initial accent or Polish with accent on the penultimate, the accentuated syllable represents a prominent feature of the phonological structure of lexical elements and thus a clear and distinct signal for starting the search and for the successful finding of lexical elements.

The information which is still needed at this point in order to be able to reconstruct completely the utterance containing several words will be processed and gained in the next step where verification is carried out by a component called the Master. Here, accessing the remaining information in the possible phonological structure and the top-down component, at this point especially syntax, pragmatics, and semantics, the missing parts of the phonological-syntactic structure are hypothesized and built into the total structure corresponding to (parts of) the utterance. After this verification, the process of speech recognition, hopefully, will end up with the identified meaning. As can be seen in Fig. 8, the Master has access to the linguistic constraints and the knowledge which, in turn, have access to the three lower levels. For the Master there is also a feed-back channel to the possible phonological structure which, again in turn, feeds back to the two lower levels. Thus it becomes quite clear that the top-down information is available to different and rather low levels of processing in speech recognition. It becomes also clear that, due to this fact, the speech signal need not be clear and distinct at every point in time. Of course, the more distinct the signal is, the easier and faster the lexical search can be because almost no support by the top-down component and no feeding-back is needed in this case. If the verification of some chosen lexical elements by the Master as to their linguistic and pragmatic correctness and of their semantic credibility comes out negative, the feed-back channel to the possible phonological structure, the hypothesized phonological structure and, if necessary, to the acoustic-phonetic basic information will be activated. Then a change of the phonological structure already arrived at will be enforced by starting the searching process anew which, finally, will arrive at an acceptable result after having passed through a number of stages a second and maybe a third time.

In this interactive process of speech recognition, it is

obvious that prosody, especially word accent, plays a direct and guiding part. Searching for lexical elements stored in LTM takes place not by using words with clearly defined boundaries but rather by using prosodic features where word accent and phrase accent or focus distinctly point to the most important semantic elements of an utterance. The syllables which are prominent due to word accent represent reliable islands in the stream of sounds and there they function as the anchor or fixation points of speech recognition. Therefore it is easily understood that word boundaries are not an absolute and significant support or even a precondition for speech recognition. Phrase boundaries, however, play an important part in dividing the speech chain into appropriate processing units. It is interesting to notice in this respect that phrase boundaries are clearly marked, often by several prosodic means. In contrast, word boundaries, are not marked in any special way. Even where morphological word structure is concerned, unstressed syllables, especially at the end of a word, as markers of concord, normally contain linguistic information which can easily be derived. Therefore it is not astonishing to learn that speech recognition systems cannot find words in the signal of continuous speech if the word, even in longer texts, are not pronounced in a staccato way, i.e. surrounded by pauses. In the speech signal there are no word boundaries but acoustically more distinct and elaborated chunks of the size of a syllable, namely the prominent and accentuated syllables.

The model of speech perception outlined here differs from previous models in several respects, although some parts, especially at the more peripheral levels, coincide. In the present model, prosodic information in the signal and in the linguistic constraints applying to different levels and structures play a leading and guiding part in solving the task of searching for a lexical element, namely the finding and identifying of, above all, basic semantic elements, making up the skeleton of meaning.

In contrast to the cohort theory, there is no activating of groups of possible word candidates all of them beginning with the same sound and the number of which will be gradually decreased as a consequence of acoustic information arriving later and of contextual constraints until, in the end, only one candidate will hold the floor. In my model, the spectral information of phonemes does not play a predominant part. Guided by the prosodic information pointing especially to the clearly marked accentuated syllable, one or more possible phonological structures not exactly defined by word

boundaries, may start for the search of lexical elements. Very often they may even act as competitors (cf. Bannert 1980).

Rather as an amendment to the Phonetic Refinement Theory, in my model the strong part of prosody in finding the most significant and central elements of meaning is duly recognized. The process of speech recognition obeys the principle of clarity. The accent pattern, prominent in the signal and easily to be discovered and processed, forms a linguistic frame or skeleton which the spectral features are subordinated to and built into. Every part of the phonological structure which is missing or indistinct, if possible, will be restored or corrected later in the interactive processes.

Another virtue of this model lies in the fact that it is applicable to the whole range of different conditions of the speech signal in verbal communication and the bottom-up component of speech perception. The top-down component is always at work. It is obvious that a distinct and good speech signal makes speech recognition easier, faster, and accurate. If the speech signal is deviant with respect to a given (band of) norm or distorted by external sources, a larger period of time will be needed in order to identify a meaning because a larger burden is put onto all kinds of memory, information paths, and feed-back channels. An increased activation of search processes and memories explains the fatigue experienced by listeners who are exposed to speech in noisy environments or to strong foreign accent for longer stretches of time.

In conclusion, then, this model also covers speech recognition under different conditions: the optimal speech signal, spoken distinctly and free from external acoustic distortions, the speaker and listener using approximately the same standard of pronunciation; the indistinct pronunciation due to lax or fast articulation; the acoustically distorted signal; the perception of the hard of hearing and the deaf; the perception under inattentiveness and non-listening of the intended listener; the geographical, dialectal, social, and individual varieties of a language; the foreign accent.

### **Recognizing foreign accent**

There is clearly no doubt that the speech signal containing foreign accent is analysed auditorily and acoustically in the



same way as the speech signal derived from standard language. First differences are to be found at the point of the acoustic-phonetic analysis. Searching for lexical elements cannot be done in real time, because the incomplete and fragmentary basic acoustic-phonetic information does not permit generating a hypothesized phonological structure leading to a possible phonological structure. As a consequence of this failure, information has to be kept in the short-term memory which puts an extra load on it, while the searching for a word is expanded by waiting for more phonetic bottom-up information and by switching on the top-down restoration and corrections components. This, in turn, will put even more strain on the recognition processes.

Another problem for lexical search arises when the possible linguistic structure points to the wrong lexical element. This is the case when a word pronounced deviatingly coincides with a different, existing word; for instance when the phoneme /y/ is rendered as the phoneme /i/ ( Swedish byta - bita 'change - bite'). In this case, the lexical search seemingly will succeed in identifying a word and finding a meaning. However, this mistake will be discovered when the word is put into the phrase or sentence where the context discloses that the wrong word was picked. The interpretation of the whole phrase or sentence has to be rejected at this stage and a new recognition process has to be started, now also by activating the restoration component. Again a greater strain is put on the processing of the speech signal. Furthermore, it has to be pointed out that the speech signal, while repetitions and retentions in the short-term memory are in full progress, continues to enter the ear, and the peripheral automatic acoustic analysis must continue its work without interruption.

The decoding processes for foreign accent should show the heaviest strain with listeners who are not accustomed to this phonological variation and who are not motivated to do such extra labour. The decoding processes for foreign accent should show the lightest strain with listeners who have developed in their long-term memory a rich component of correction rules for foreign accent - which is closely related to the typical features of foreign accent of a given L1 - and who really want to understand foreigners by activating both the feed-back path (the correction component) and the access path of top-down information (the restoration component).

## FOOTNOTES

- <\*> I would like to thank Klaus-Jürgen Engelberg, David House, Bernhard Keck, Gerhard Rigoll, and Herbert Tropic for helpful and valuable comments and contributions to this paper.  
This research was supported by the Bank of Sweden Tercentenary Foundation.
- <1> The manipulations of the speech signal were made at the Department of Linguistics, Uppsala University. I am very grateful to Sven Öhman for his kind support and Lennart Nordstrand for his expert assistance.
- <2> It became evident from preliminary tests using filtering as a means of distorting the acoustic speech signal that intelligibility would not be decreased to a sufficient and desired degree. Therefore the planned experiment with filtered speech was excluded (cf. Bannert 1984).
- <3> The listening tests were disguised as reaction tests attempting to measure the listeners' ability, as quickly as possible, to prompt the utterances spoken by foreigners and presented under hard listening conditions.
- <4> One utterance (L1 = Greek) that was part of the test was excluded from the presentation of the results as the manipulation of voicing and voicelessness of the obstruent cluster by LPC-synthesis is not reliable in this respect.
- <5> Quantity in Standard Swedish is manifested as the complementary length pattern /V:C/ vs /VC:/ in the stressed syllable. Phrase rhythm means the temporal relationships between successive syllables. Accent is a tonal feature of syllable prominence and is manifested as a change of  $F_0$  in or in connection with the accentuated syllable.
- <6> The parameter of volume (intensity) of syllables was not included in the manipulations of accent. This does not mean, of course, that intensity might not be a contributing factor in the complex feature of accent. It is believed, however, that intensity is not an essential feature of normal word accent (as opposed to contrast or emphasis).
- <7> These prosodic deteriorations corresponding to the

features of phrase rhythm, quantity, and pitch accent, had a detrimental effect on the identity of the utterances. In several instances of demonstrations where the deteriorated Swedish utterances were played to linguists, phoneticians, and experienced teachers of Swedish as a second language, these disguised utterances were accepted as foreign accent and associated with certain first languages L1.

- <8> As individuals, listeners react differently to the presentation of the stimuli. While some of them always try to respond even guessing to some degree, others hesitate to respond at all if they are not quite sure about the intended structure.
- <9> For these larger units no definition is provided here. Yet it is assumed that the notion is well-established.
- <10> Some possible elaborations in certain respects may look like parts of the model in Lea et al. (1975).
- <11> The dimensions of voice quality and volume will also be analysed on this level. These processes are only mentioned to complete the picture and will not be dealt with here.

## REFERENCES

- Bannert R. 1975. The significance of vowel features in the perception of complementary length in Central Bavarian. Lund University, Department of General Linguistics and Phonetics, Working Papers 12, 1-45.
- Bannert R. 1978. Prosodiska egenskapers effekt på förståeligheten. In: Kommunikativ kompetens och fackspråk. Linnarud M. and Svartvik J. (eds.). Lund
- Bannert R. 1980. Phonological Strategies in the Second Language Learning of Swedish Prosody. PHONOLOGICA 1980, 29-33. Innsbruck
- Bannert R. 1984. Prosody and intelligibility of Swedish spoken with a foreign accent. Nordic Prosody III, C-C.Elert, Johansson, and E.Strangert (eds.). Acta Universitatis Umensis, Umeå Studies in the Humanities

- Bannert R., Engstrand O., Eriksson H., and Nordstrand L. 1982. Phonetics and Identity. Uppsala University, Department of Linguistics, Report 8, 22-58
- Bruce G. 1977. Swedish Word Accents in Sentence Perspective. *Travaux de l'Institut de Linguistique de Lund XII*. Lund
- Cole R.A. and Jakimik J. 1980. A Model of Speech Perception. In: *Perception and Production of Fluent Speech*, 133-163. Cole R.A. (ed.). Hillsdale
- Elman J.L. and McClelland J.L. 1984. Speech Perception as a Cognitive Process: The Interactive Activation Model. In: *Speech and Language: Advances in Basic Research and Practice*, Vol. 10, 337-374. Lass N.J. (ed.). New York.
- Elman J.L. and McClelland J.L. 1985. Exploiting Lawful Variability in the Speech Wave. In: *Variability and Invariance of Speech Processes*, Perkell J.S. and Klatt D.H. (eds.). Hillsdale, N.J.
- Forster K.I. 1979. Levels of Processing and the Structure of the Language Processor. In: *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*, Cooper W.E. and Walker E.T. (eds.). Hillsdale, N.J.
- Gårding E. and Bruce G. 1981. A Presentation of the Lund Model for Swedish Intonation. Lund University, Department of Linguistics and Phonetics, Working Papers 21, 69-75
- House D. 1985. Sentence prosody and syntax in speech perception. Lund University, Department of Linguistics and Phonetics. Working Papers 28, 91-107
- Klatt D.H. 1979. Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access. *J. of Phonetics* 7, 279-312.
- Klatt D.H. 1980. Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access. In: *Perception and Production of Fluent Speech*, 242-288. Cole R.A. (ed.). Hillsdale, N.J.
- Lea W.A. 1980. Prosodic Aids to Speech Recognition. In: *Trends in Speech Recognition*, 166-205. W.A. Lea (ed).

Englewood Cliffs, N.J.

- Lea W.A., Medress m.F., and Skinner T.E. 1975. A Prosodically Guided Speech Understanding Strategy. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, 30-38
- Maassen B. and Povel D.-J. 1985. The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech. JASA 78, 877-886
- Marslen-Wilson W.D. and Welsh A. 1978. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology 10, 29-63
- Morton J. 1979. Word Recognition. In: Psycholinguistics 2: Structures and Processes, 107-156. Morton J. and Marshall J.C. (eds.). Cambridge, MIT Press.
- Nespor M. and Vogel I. 1983. Prosodic Structure Above the Word. In: Cutler A. and Ladd D.R. (eds), Prosody: Models and Measurements, 123-140 (=Language and Communication 14). Berlin
- Nooteboom S.G. 1981. Lexical retrieval from fragments of spoken words: Beginnings vs. endings. Journal of Phonetics 9, 407-424
- Osberger M.J. and Levitt H. 1978. The effect of time errors on the intelligibility of deaf children's speech. JASA 66, 1316-1324
- Pisoni D.B. 1984. Acoustic-Phonetic Representations in Word Recognition. Indiana University. Research on Speech Perception, Progress Report No.10, 129-152
- Pisoni D.B., Nusbaum H.C., Luce P.A., and Slowiaczek L.M. 1985. Speech Perception, Word Recognition and the Structure of the Lexicon. Speech Communication 4, 75-95
- Sendlmeier W.F. 1985. Psychophonetische Aspekte der Wortwahrnehmung (= Forum Phonetikum 35). Hamburg
- Shipman D.W. and Zue V.W. 1982. Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems. ICASSP 82, 546-549.

Svensson S-G. 1974. Prosody and Grammar in Speech Perception.  
University of Stockholm, Monographs from the  
Institute of Linguistics 2

van Wijk C. and Kempen G. 1984. From sentence structure to  
intonation contour. Germanistische Linguistik 79-80,  
157-182. Müller B.S. (ed.). Hildesheim