FROM PROMINENT SYLLABLES TO A SKELETON OF MEANING: A MODEL OF PROSODICALLY GUIDED SPEECH RECOGNITION

ROBERT BANNERT

Department of Linguistics and Phonetics, Lund University, Lund, Sweden

ABSTRACT

A model of speech recognition is sketched where the guiding role of prosody, especially the pathbreaking function of the accent syllables, is duly stressed. The relationships between the accent syllables and the root syllables of words provide the listener with a skeleton of meaning which will be completed and, if necessary, restored in further stages of the recognition processes. In a hierarchical organization of linguistic structures and processing levels, information flows between the acoustic-phonetic and the semantic level in a purposeful and optimal way interacting with phonological, morphological, syntactic, and pragmatic information.

INTRODUCTION

For quite a long time, speech perception and speech recognition has challenged the mind and skill of students of various fields of research like psychology, linguistics, phonetics, and engineering. In spite of all the enormous progress that can be witnessed, we have to state today that the problems of recognizing fluent speech, spoken by different speakers, are far from being solved as far as the fundamental principles characteristically employed by human listeners are concerned. The explanation for this state of the art has to be sought above all in our insufficient knowledge of the processes leading from the acoustic signal to the understanding of meaning conveyed by the speech signal.

In recent years, the significance of prosody in speech recognition has been recognized to an increasing degree [e.g. 1, 2]. The present paper is intended to contribute to a better understanding of the processes involved in speech recognition. Experimental data point to the important role, in relation to their acoustic and semantic features, that syllables made prominent by word accent play in the processing of the speech signal by the listener.

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), and their prosodic structure, e.g. 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. 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.

EXPERIMENTAL DATA

In a series of experiments, samples of Swedish, spoken with a strong foreign accent and deviating with respect to the distribution of word accent, were corrected temporally and tonally and thereafter presented to native Swedish listeners under various hard listening conditions. Deviant speech aggravates speech perception because the acoustic information contained in the speech signal and constituting the initial information to the processes of speech recognition may differ markedly from the normal and expected standard of pronunciation. Some interesting results emerged from manipulating certain features of the speech signal in a controlled manner by means of LPC-speech synthesis of high quality and then studying listeners' reactions to the manipulations. A detailed description of the method used and the results are given in [3, 4].

Evaluating listeners' responses to utterances manipulated in this manner, one observation is prevalent: It always seems to be the accent pattern that is picked up by the listeners. Accent pattern means the linear succession of accented and unaccented syllables in an utterance. The same accent pattern is to be found in the listeners' response, even if the accent pattern is incorrect in the stimulus, although the response differs from the intended utterance with respect to its semantic, syntactic, morphological, and phonological structure. If by way of speech synthesis the wrongly positioned word accent is moved to the correct syllable, listeners' responses will change in the same way. A typical example of changes of accent pattern is illustrated in the follow- ing (the position of word accent = primary stress is indicated by the symbol '): The Swedish phrase "i 'samhā:llet" (in society) showed the wrong prosodic structure "i sam'he:let" in the devi- ant foreign accent rendering. This stimulus was heard as "i sin 'he:lhet" (on the whole) or "utan 'te:ve" (without TV) by listeners who obviously focussed on the accent pattern or the distribution of the word accent. However, when the tonal move- ment, representing the most essential feature of word accent, is moved from the wrong syllable "-'he:-" to the correct syllable "'sam-", the pattern of listeners' responses is changed in accor- dance with the correct accent pattern. Listeners now heard "i 'sandträ:det" (in the sand tree), "i 'samlingen" (in the collection) or "i "handlingen" (in the action), all of them showing the ident- ical distribution of word accent on the second syllable of the stimulus. The number of syllables in the listeners' responses was always identical. At the same time, it has to be noted, of course, that the accent syllables of the listeners' responses share a cer- tain amount of spectral features with the accent syllable of the stimulus.

OUTLINE OF THE MODEL

Assuming certain general principles in some existing models of word and speech recognition (e.g. [5]), the most relevant features of a prosodically guided model of speech recognition will be outlined. A more detailed description is to be found in [4]. The model is summarized graphically in Fig. 1.

The hypothesized phonological structure resulting from the acoustic-phonetic analysis of the speech signal and the restructuring working of phonetic, phonological, and morphological knowledge is not totally specified. The possible phonological structure that acts as the search unit for lexical items is assumed to be an accent group, i.e. a chunk of a linguistic structure containing as its kernel the accent syllable surrounded by other unstressed syllables. No word boundaries are marked nor needed.

All information of linguistic and pragmatic kind may be used by the various stages and processes of speech recognition at all times and wherever necessary and useful. A close and optimal acting together of bottom-up and top-down information even at low levels where the first linguistic interpretation of auditive-acoustic information occurs and non-linguistic short cuts bypassing all the hierarchically structured acoustic and linguistic levels, are assumed for speech recognition.

The acoustic analysis of the speech signal is performed in two different channels, i.e. the prosodic and the spectral one (cf. [1]). Quite often the auditive-acoustic analysis cannot always result in a complete phonetic basic structure due to acoustic distortions from outside and assimilations and reductions in the signal itself.

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. [5]). The phonetic interpretation provides the basis for the acoustic-phonetic basic information about the chunk of the speech signal to be processed.

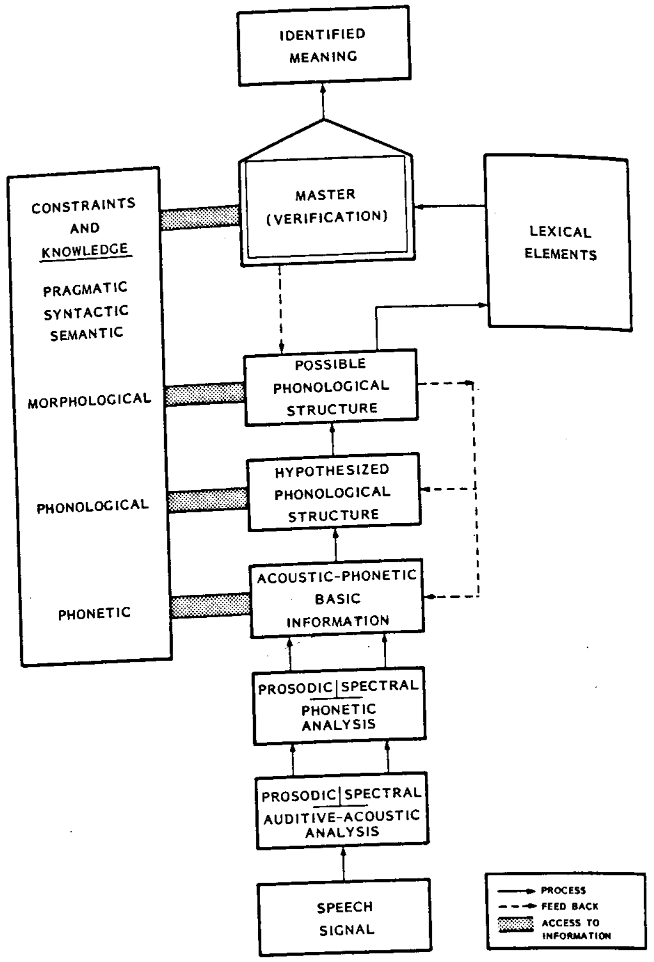

Fig. 1 A model of prosodically guided speech recognition

The acoustic-phonetic basic information is structured according to prosodic and spectral features. The prosodic features provide the position of the accented 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 accented 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. 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.

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 semantic elements of the lexicon are arranged in a multidimensional 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 accented syllable as its most important search criterion. Therefore it can be assumed that the search starts out for phonological representations of lexical elements showing the identical accent pattern and some 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.

A verification component, the Master, has access to the linguistic constraints and the knowledge which, in turn, have access to the 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 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 credability 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 the long-term memory 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 a 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 words, 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 accented syllables.

The model of speech perception outlined here differs from previous models in several respects. 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 accented 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. [6]).

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 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.

REFERENCES

[1] Svensson S.-G. 1974. Prosody and Grammar in Speech Perception. University of Stockholm. Monograph from the Institute of Linguistics. No. 2

[2] 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

[3] Bannert R. 1984. Prosody and Intelligibility of Swedish Spoken with a Foreign Accent. In: Nordic Prosody III, C.-C. Elert, I. Johansson, and E. Strangert (eds). Acta Universitatis Umensis, Umeå Studies in the Humanities 59, 7-18

[4] Bannert R. 1986. From prominent syllables to a skeleton of meaning: A model of prosodically guided speech recognition. Lund University, Department of Linguistics and Phonetics, Working Papers 29, 1-30

[5] Pisoni D.B. 1984. Acoustic-Phonetic Representations in Word Recognition. Indiana University. Research on Speech Perception. Progress Report No. 10, 129-152

[6] Bannert R. 1980. Phonological Strategies in the Second Language Learning of Swedish Prosody. PHONOLOGICA 1980, 29-33. Innsbruck