TEXT-GUIDED AUTOMATIC ANALYSIS OF THE SPEECH SIGNAL AS A POSSIBLE MEANS OF APPROXIMATING AUTOMATIC SPEECH RECOGNITION

HELMUT RICHTER

The present paper deals with some of the differences between and common aspects of automatic speech recognition (ASR) and instrumental phonetics in a more traditional understanding. Let me start with a series of assumptions and definitions which I feel to be adequate to the somewhat dialectic development of the argument.

(A1) It is possible to represent speech events or particular aspects of them by the speech signal. It is also possible for a human observer to represent speech events or particular aspects of them symbolically by text. I will call any text obtained from human observers such that the outcome is accepted by a relevant section of the speech community ORDINARY TEXT. Symbolic text is to consist of tokens, the single type of which from the intended point of view may be called a DISCRIMINATE.

(A2) One useful representation, in symbolic form, of speech events will be characterized by its discriminates being uniquely composed out of a restricted inventory of other symbolic units; let me label those smaller units DISCRIMINATORS.

One can conceive of written words as discriminates and of graphemes as discriminators, but also of written phonemes as discriminates and something like the graphic representation of their distinctive features as discriminators; it is the relative location of the discriminate-discriminator step which I want to lay stress upon. Similarly I am not primarily concerned with whether either type of units involves a sign value.

(A3) It is sometimes possible to segment a speech signal according to the sequence of discriminates in a text based upon the same speech event. In that case, the text units obtain a specific descriptive relevance by being instances of co-ordinating two forms of representing the speech event, which are not necessarily linked in our assumptions and definitions. I will make reference to this peculiar status by the word S-DISCRIMINATE, while a segment (or segment class, respectively) being the correspondent of a s-discriminate can be labelled D-SEGMENT. (Note that the formulation is open for a variety of uniqueness postulates, one of which may be selected *ad hoc.*)

Given the assumptions (A1)-(A3), it follows:

(P1) The discriminators of s-discriminates are not by necessity S-DISCRIMINATORS; that there are *d*-segments in the speech signal is no guarantee for that also DD-SEGMENTS can be found in the same signal relative to the same symbolic text. (The new terms are self-explanatory expansions of the terminology from [A3].)

(A4) Given one or more signal representations of the speech event, a symbolic representation of that event can be obtained automatically. It will in the long run be even possible automatically to produce text which comes reasonably close to ordinary text. (A4) may be regarded as a definition of AUTOMATIC SPEECH RECOGNITION.

(A5) ASR will involve processes tending to uniquely build up relevant states out of a restricted set of elementary states; these latter are referred to in what follows as COMPONENTS IN ASR. (Note a parallelism between (A5) and the definition of discriminators under (A2).)

Taking also (A4) and (A5) into account, it can be stated:

(P2) There must be one level, where ASR arrives at *d*-segments being equivalent to the discriminates of ordinary text (e.g., words). As soon as it is possible automatically to infer, from these discriminates, the discriminators commonly used (e.g., letters), then no co-ordination or mapping will be necessary between the components in ASR and the discriminators of ordinary text. Since phonemic text was not awarded a special category of ordinary text, this is also to mean that the components in ASR are not necessarily equivalent to phonemes or distinctive features.

So far, what has been said can be understood as an argument for the emancipation of ASR from phonemics. Automatic speech recognition will scarcely turn out to be automatic phoneme or sound recognition. The 'auteme' (automata-phoneme) conception followed in the project on ASR at the Institut für Kommunikationsforschung und Phonetik at Bonn is, in this sense, an emancipated one (see e.g., Tillman 1967). Beyond that, the argument could easily be extended such as to formulate a related scepticism concerning the mutual mappability or invariance of the segmental units which can be distinguished in the articulatory, acoustical, and auditory manifestations of the speech event, i.e., in more traditional terms a scepticism as to the validity of a 'transposition' postulate for these units (Richter 1967).

Instrumental phonetics should, however, not be restricted to ASR. The question of how linguistically established sound units are 'realized' in the speech signal remains a valid one, PROVIDED THAT the linguistic establishment of sound units proves in itself sensible for socially relevant ends (as phonemics did for 'reducing language to writing'), and provided that the speech signal must be taken into consideration at all when one specific end is pursued with the aid of linguistically established sound units. (How to conceive of Applied Linguistics under such premises, has been pointed out by Kohler, Tillmann and Richter on various occasions (see e.g., Kohler 1970, Tillmann, and Richter, in press).

It is obvious that getting signal correlates of linguistically established sound units and their retrieval would be considerably facilitated if one disposed of an automatic procedure to co-ordinate the acoustical speech signal with given phonetic text. By this it is roughly explained what is meant by TEXT-GUIDED AUTOMATIC ANALYSIS (TAA) in the paper's heading. More technically, one can define this type of analysis as a method of automatically obtaining d-segments or dd-segments related to units in an ordinary text given in advance, which is supposed to contain s-determinates or s-determinators, respectively. In practice this can be very difficult (indirect evidence has been provided by phonometry), and yet is far from being ASR.

Nonetheless TAA could indirectly further the specific ends of ASR. In other words, developing the method seems to be not only a desideratum of instrumental phonetics with linguistic orientation, but additionally motivated by some practical or heuristic needs of ASR. In order to dispose of a strategy for combined efforts in both areas, an elaborate theory will be necessary. It appears that this must be a theory posing the problem under the angle of adaptive information retrieval systems.

As far as I can see, this particular theory has not yet been developed. I would, however, venture certain hypotheses about why TAA according to a given text could function as an approximation of ASR. There are, I think, two main clusters of pertaining expectations:

(H1) There can exist partial mapping relations between the *d*- and *dd*-segments for ordinary text (which is built up out of *s*-determinates and *s*-determinators) and the components in ASR, even if there is no equivalence. The partial mapping may concern either subsets of the inventories involved and/or restricted aspects of discrimination pertaining, however, to key subsets in the inventories.

(H2) Even where there are no such favourable 'material' conditions like those indicated under (H1), similarities or invariances of procedure between TAA and ASR can play an important rôle. TAA thereby could provide for essential insights and concrete experience as to what kind of adaptive mechanisms are to be used in ASR. Among the similarities in question I would mainly list:

(a) variations in signal detection and signal evaluation controlled by the respective point in the text and/or some preliminary result of the analysis;

(b) comparative and iterative operations as prerequisites for reducing uncertainties by inference.

Institut für Kommunikationsforschung und Phonetik
Universität Bonn

REFERENCES

Kohler, K.
1970 "On the Adequacy of Phonological Theories for Contrastive Studies"; in *Contrastive Linguistics*, G. Nickel, ed. (Cambridge).

Richter, H.
1967 "Die Zweistufentheorie Šaumjans und das phonometrische System der Varianten", *Phonetica* 16:156-184.

Tillmann, H.G.
1967 "Akustische Phonetik und linguistische Akustik", *Phonetica* 16:143-155.

Tillmann, G. and H. Richter
1972 "Zwei Beiträge zum Problem der anwendungsorientierten phonetischen Analyse", *International Review of Applied Linguistics and Language Teaching*, Sonderband gal '70:223-235.

DISCUSSION

IIVONEN (Oulu)
As far as I understand, your work is of a theoretical nature. Do you think that one could say — on a purely theoretical basis — if the automatic recognition of speech is possible or not without working with some hardware or other empirical methods?

RICHTER
From a theoretical point of view, I believe it is a reasonable assumption that ASR is possible; maybe one will practically only succeed in sensible approximations. But this is not the only consideration that counts. What is needed in order to prevent phonetic research from being anecdotal, are clear-cut aims. ASR is to provide us with one such aim and is thus valuable from a science-logics point of view.

PADDOCK (Wolfville, N.S.)
Do you expect that the 'machinery' which would actually carry out ASR will need to have access to that kind of information about lexicon, syntax, or semantics which a human being actually uses in his recognition of speech.?

RICHTER
As to empirical work, I can refer to the papers that were read about detail investigations by my colleagues from Bonn.

PADDOCK
It would seem that at some point in the ASR process one must convert or transform physical measurements into some kind of linguistic primitives. Do you feel that this transformation is the major problem facing ASR?

RICHTER
One must distinguish between a 'communicative' make-up of analytic processes and everyday communication, the latter being the object of the former. (Ungeheuer's distinction between the communicative and the extracommunicative might be referred to in this connection.) It is obvious, however, that communicatively made up ASR will include the use of higher type information by the automatic process (hypothesis formation about what really MAY have been said, etc.).

TILLMANN (Bonn)
I would like to know if your prior experience with phonometry encouraged you to postulate TAA.

RICHTER
In a sense; it can be rather discouraging to do phonometrical work without disposing of automatic segmentation procedures. So TAA would be a chance to accomplish phonometry.

SOVIJÄRVI (Helsinki)
I would like to add a little comment concerning the large co-operation and teamwork you need when you develop your important work. In Bonn you have a good opportunity to get many kinds of help from your colleagues who represent different, relatively special, areas concerning automatic speech recognition.

RICHTER
This is certainly true. The point which seems, in my paper, most pertaining to your remark is concerning the necessity of adaptive information retrieval systems for ASR. In this respect, phonetic work done at our institute tends to converge with the work which is e.g., undertaken at the institute in the field of linguistic data processing. On the other hand, I just tried to indicate how linguistically oriented phonetics too could be helpful for ASR.