APPENDIX

VOICE RECOGNITION BY MAN, ANIMAL, AND MACHINE

HENRY M. TRUBY

DISCUSSION

FRY (London)

Those of us who have known Professor Truby for some long time are well aware that when we hear a paper from him, we are going to hear something original and indeed highly individual. In this respect, we have certainly not been disappointed today. On this occasion, he has in addition ranged over many different topics so that in this respect it is not possible to do justice to his presentation in discussion and I shall therefore confine myself to just one point, the identification of individual speakers, and try briefly to put forward some evidence which I feel sure will be of interest to Professor Truby, and I hope to others.

To some extent, I think I must join issue with him when he says that the differences between speakers are carried by idiosyncratic physical features which are not apparent to even the most highly 'trained' ear. The fact is that all of us are able to recognize voices which are very familiar to us and to do so with a high degree of certainty. In this respect then, we all have highly trained ears and we are able to pick up the physical features which distinguish one speaker from another. The situation here is one which occurs frequently in the whole field of speech and language: speaker identification is an operation which the listener can carry out and the task facing us professionally is to specify how this is done. Our efforts in this direction have not so far been outstandingly successful but there are at least good indications of the general areas which need to be explored.

There are broadly three aspects of an individual's speech which will help a listener to identify him. First his pronunciation may have some idiosyncratic features, that is to say, there may be things in his articulation and particularly in the timing of articulatory movements which mark him off from other speakers with the same regional and social dialect. Second, there may be certain properties of his vocal tract and of the way in which he uses it which may help to identify him; these features are quite likely to register acoustically in the region of the higher formants, the fourth, fifth and so on. Third, there will very probably be features in his phonatory activity which characterize his speech as an individual. It is about this last aspect that I want to make some brief remarks.

There is already a body of evidence that the larynx output, the source function, plays an important part in speaker identification. The fact that Vocoder systems have generally failed to find acceptance for ordinary telephone use is attributed to the lack of naturalness and the difficulty of speaker identification which their use entails; this negative evidence is supplemented in a positive way by the observation that the Vocoder becomes much more acceptable if it transmits all the available information about the larynx output.

The laryngograph, which many of you have seen demonstrated at this Congress, provides a good means of adding to the evidence in this area and some pilot experiments have already been begun in London. A group of twelve speakers who were very familiar with the voice of everyone else in the group were asked to record three short English sentences in which not only the words but also the intonation patterns were specified. Tape-recordings were made of both the output from the mouth and from the laryngograph electrodes. The point of interest is that the twelve subjects, when allowed repeated hearings of the laryngograph recordings only, without any information as to the correctness of their identifications, were soon able to do far better than chance in identifying all the speakers in the group; in fact, one member was able to identify all twelve correctly simply on the basis of the larynx output, without any vocal tract information. In an extension of this pilot experiment, the same speakers recorded three vowel utterances, /i/, /a/ and a central vowel, on a monotone in isolation, both voiced and whispered. They were then asked to identify all the speakers when supplied with the recording of the mouth output of voice and whisper and the laryngograph output of the voiced utterances, each separately. These three conditions represent source and vocal tract information combined, source information alone and vocal tract information alone. In the first condition the subjects scored a very high proportion of correct identifications, as one would expect, but it is significant that they in fact returned higher scores when they received source information alone than when they were supplied with vocal tract information alone.

As I have said, this work is at a very preliminary stage but the results do once again point to characteristics of the larynx output as having probably considerable weight in speaker identification. This does seem to be one of the directions which it may be profitable to pursue in our study of speaker identification and in our attempts to answer at least some of the many questions raised by Professor Truby this morning in his highly interesting and informative paper.

TRUBY

Here, as on many previous occasions, I find the exceptional fluency and considerable, many-faceted experience of Professor Fry catapulting him magnificently into mid-critique with no time wasted, and I consider myself privileged at being the focal point of his attention on this occasion.

"To some extent" and "I think", with which he tempers his "issue joining", are not only a reflection of his habitual diplomacy, politeness, and carefulness, but, in this instance, serve as the bridgings which unite his and my independently nurtured evaluations of those "idiosyncratic features" on the basis of which a given individual successfully identifies another individual (i.e., discriminates individuals) on the so-called oral-aural plane. Dr. Fry and I both well know that such "identifications" ARE commonplace — with high levels of consistency, regularity, reliability, validity, etc. — BUT that such expected recognition is punctuated, here and there, with instances of dismal, auto-astonishing failure, just as regularly occurs in VISUAL-recognition experience. All of us have at some time or other mis-identified a perfect stranger as an acquaintance or even as a close friend, either auditorily — especially over the telephone — or visually, to our great astonishment at the moment. Thus it is that whether "even the most highly 'trained' ear" is fooled or not, the fact remains that CERTAIN idiosyncrasies of VOICE (as well as of idiolect, and thus of speech) present in the acoustic signal are NOT apparent to "the naked ear", even though, often, enough details ARE apparent to permit the familiar speaker recognition with which we live dailily. In a sense, "the ear" makes an identifying CARICATURE of the distinguishing features manifested in overabounding detail on sound spectrograms.

I should also like to expand on the classification and enumeration cited by Professor Fry re "individual speech aspects": As I see it, ALL SPEECH PERFORMANCE IS IDIOSYNCRATIC — the idiosyncrasy dichotomy under examination in my paper being resolved as IDIOLECTAL versus *non*-IDIOLECTAL; accordingly, "pronunciation" is essentially idiolectal, "timing" is essentially idiolectal, "vocal tract properties", are NON-idiolectal, "the way in which he uses it" is ambiguous and thus idiolectal in one sense and NON-idiolectal in another sense (this being one of the points of those portions of my paper treating "voiceprinting" *per se*), and "features in his phonatory activity which characterise his speech as an individual "are basically NON-idiolectal. [With all due respect for Dr. Fry's opinion, I myself find that neither idiolectal nor NON-idiolectal features are necessarily more likely than not "to register acoustically in the region of the higher formants", since all manifestations of idiosyncrasy (idiolectal and non-idiolectal) implicate the ENTIRE speech- and voice-relevant acoustic spectrum, its variations in time, and all inherent and conditioned hiatuses. It is certainly true that all inter-speaker variations are manifested the most eye-catchingly in the higher-formant regions, BUT, so are all INTRA-speaker variations! And there lies the rub... as a result of which one must tread very carefully, keeping in mind that visible-acoustic details can be expensively disarming.] Thus, assignment within the dichotomy IDIOLECTAL/NON-IDIOLECTAL is clearly dependent upon the assigner's comprehension — and thus, definition — of "pronunciation", "timing", "features", etc., hence my precautionary modifiers and clarifiers "essentially", "ambiguous", and "basically".

Most comforting of all, a careful scrutiny of Dr. Fry's account of the "laryngo-

graph" and related "pilot experiments" reveals his confirmation that "source function" (i.e., "larynx output") information appears to be critical in "speaker identification". THIS parameter is clearly indicative of idiosyncrasy without reference to IDIOLECT, and as I indicated in my paper, that which is generally touted as "voice-printing" is, in truth, "speechprinting", the actual VOICE being a reflection of laryngal and even supraglottal participations individualistic to the exclusion of idiolectal considerations. That the supraglottal idiosyncrasies have some significance is borne out in the *whisper* evaluations, such features, however, being never than minimally reliable.

I do thank Professor Fry for his complimentary critique and sincerely regret that time did not permit him to address himself to other of the "many different topics" I felt were relevant on this occasion to a discussion of Voice Recognition, HOWEVER implicating man, animal, or machine.

MOL (Amsterdam)

A very important aspect of Dr. Truby's paper is what we might call with a big word 'forensic phonetics'.

My own attitude towards forensic phonetics has always been one of severe scepticism, though I am not a stranger in this field.

I don't think that voice identification by ear is impossible, although I should call this procedure an art instead of a science. I am inclined to call the interpretation of objective measurements for identification likewise an art and not yet a science. According to Dr. Truby, this interpretation should be in, what he calls, the proper hands.

Before I came here, I toyed with the idea that the results of forensic phonetics could only be used to scare a criminal into a confession. After hearing Dr. Truby's excellent paper in which he presents his own views and experience, I am inclined to be a little bit more optimistic from now on. I agree with Dr. Truby that modern instrumental methods will provide us with many acoustic cues upon which we may base our conclusions. I still fear, however, that the number of proper hands into which the interpretation of identification measurements may be confided is, at least at the moment, rather limited. I am convinced however, that Dr. Truby's hands are sufficiently proper to help us put the problems of voice identification on a more scientific basis.

TRUBY

I have long admired the succinctness with which my respected colleague Professor Hendrik Mol is able to evaluate cruxes and other troublesome sectors of hypothesis and practicum. Thus I am especially proud to be the recipient of professional compliments from him.

If, for instance, my "hands" are indeed among those "proper" for "help[ing] us put the problems of voice identification on a more scientific basis", I am anxious to place those hands constructively at the service of a Voice Recognition Commission (which I propose) with final-decision and other appellate jurisdiction in all Voice Recognition matters and instances. Only thus can "voice-printing" be treated "on a more scientific basis", as the relevant archives are accumulated which will put "voiceprinting" on a reliability par with 'fingerprinting', 'palmprinting', 'footprinting', and other accepted dermatoglyphic procedures, whether for criminological, diagnostic, or simple identificatory application. [And see my reply to the critique of Adrian Fourcin.] It should be borne in mind that the first definitive report, *Finger Prints*, was published a mere eighty years ago, and that fingerprints had been admitted as substantiating evidence in only one state in a single instance at the turn of this Century! And much of the same criticism presently being directed at "voiceprinting" was vociferously put forth — by the usual few — against fingerprinting and its implications variously.

The protection of the public is alone enough to make my proposal FOR the creation of such a Commission unopposable, especially in light of the present lack of knowledge ABOUT Voice Recognition which has surfaced in criticisms of and/or applications of Voice Recognition procedure. For instance, in forensic considerations especially, it is not enough simply to understand the design, physical nature, and PHYSICAL-NATURE FUNCTION of acoustic-measurement instrumentation; nor just to be, however thoroughly, conversant with the linguistically significant features of articulatory phonetics; nor just to have made however MATHEMATICALLY astute acoustic-phonetic observations about these linguistically-significant articulatory gestures. For valid VOICE Recognition evaluations, the evaluator must have accumulated experienced and dedicated training in the sound-spectrographic analysis of both the LINGUISTICALLY-significant and the NON-linguistically-significant acoustically manifested features of speech-sound and other voice-sound output. A VR Commission, if properly staffed and facultied, would have to acknowledge any demonstration of likeness or difference introduced by "voiceprint expertise" before that demonstration would be acceptable in forensic affairs. The Commission would not itself conduct laboratory investigation, but it would examine and evaluate any corpus of acoustigraphic analysis introduced as legal evidence.

Though I have spoken pointedly to the issue of "voice identification by ear" (see fn. 19 and related text), at least to one of the more dramatic aspects, the subject implicates too broad a range of variables for quick discussion — it is neither "art" nor "science", however, since it is not dependent on either the artistic or the scientific ability of the identifier, but on a vague relationship we can call FAMILIARITY. But "the interpretation of objective measurements for identification" is dependent solely upon the training and experience — and always on the acuity — of the interpreter in the selecting of those physical details relevant to a particular visible-acoustic identification. I'd certainly rather hang MY hopes on the sound-spectrographic analysis of a

substantial segment of speech output than on the vacilatory acoustic-memory resources of Mr., Mrs., Ms., or Miss John (or Jane) Q. Public!

As for the instrumental analysis procedure *per se*, at this particular point in history only "closed trial" investigations should be attempted in any case, i.e., given substantial recordings of ALL "suspects", an additional substantial recording of any PARTICULAR "suspect" from the group should provide positive "voiceprint" identification in "the proper hands". But NO ONE should attempt "open trial" matching, except for practice. Nor is "same-single-word comparison" either adequate or even relevant to the problem.

It is encouraging to find my friend Professor Mol encouraged, and I compliment HIM on his cautious and openminded attitude and recommendations.

PILCH (Freiburg-im-Breisgau)
Can voice characteristics be classified into a limited number of voice types? Such voice types are, I believe, needed for phonetic and phoniatric as distinct from forensic purposes.

TRUBY
As with the awareness emphasizing every thought of his *Phonemtheorie* and of his oral critique of my paper, the above written commentary of Professor Herbert Pilch is not only succinct but couched in meticulously referential terminology. For instance, the expression "voice characteristics" is punctiliously to the point, as is the reiterated expression "voice types". And equally notable is the accurate antecedentry of these expressions with "phonetic and phoniatric", to the avoidance of reference to phonemic or phonological aspects.

Yes, I believe that "voice characteristics" (as recoverable from sound spectrograms) CAN be accurately "classified into a limited number of voice types", but that, as with other sorts of pattern classification systems, open provision must be anticipated for subtypical variations bound to appear as the archives expand.

Such a classification system would predictably appear as voiceprinting comes to figure in forensic and other identificatory concerns and would in fact be indispensable to the optimal application of the procedure.

It would, in addition, being even "SUBphonetic", bring greater order TO phonetic, phoniatric, and all other linguistic descriptions, and I thank my long-time associate Dr. Pilch, for providing me the opportunity to speak to this particular aspect, as well as for his oral remarks re LINGUISTIC ANALYSES as an aspect of Voice Recognition.

FOURCIN (London)
I should like to make two points.

First, spectrograms are not to be compared with finger print patterns. The latter are formed in the foetus and suffer no important changes throughout life. Speech spectrograms, or voiceprints as they are sometimes improperly called, give speaker identity information which is not easy to employ at best and not stable with time.

Second, a particular disadvantage with spectrographic speaker identification arises when one speaker tries to imitate another. A. W. F. Huggins and I have analysed this sort of speech and we found that it presents a much more difficult problem than ordinarily occurs.

TRUBY
To answer this particular critique of my old friend and kindred laboratory spirit, Dr. Adrian Fourcin, is, unfortunately, to differ with his opinions about certain commonly attention-attracting aspects of Voice Recognition considerations.

To begin with, as I have meticulously indicated in my paper (for which, please see footnote 22 and related text), VOICE spectrograms are most EXPLICITLY comparable with fingerprint patterns. Those anatomical and neuromuscular voiceprint-pattern particulars idiocratic from individual to individual are, like fingerprint patterns, also "formed in the foetus" and do manifest individuality "throughout life", though admittedly with less inflexibility than in dermatoglyphic considerations and specifically as influenced by growth and development changes of a non-linguistically significant nature. In these regards, VOICEPRINTS is a most "proper" and apt terminology, "speech spectrograms" being indicative of only that aspect of voice idiocrasy which has LINGUISTIC relevance — the aspect which has made for a general, and almost universal, MISCOMPREHENSION of voice identification potential. As with fingerprint loops and whorls, IDIOLECTAL features contribute toward general classification and indexing, but the VOICE idiosyncrasies apparent even on sound spectrograms are analogous to such minute skin-ridge details as the differentiating bifurcations, interruptions, and terminations of fingerprint lore.

I must acknowledge Dr. Fourcin's implication that speaker identity information "is not easy to employ", largely due to the fact expressed in the text related to my footnote 18, namely that there have been no comprehensive or even systematic studies of how even the SPEECH of individuals is manifested on sound spectrograms, much the less how the individual VOICE is manifested sound-spectrographically!

I would agree with Fourcin and Huggins that certain accomplishments of speaker imitation can complicate speaker identification, but my own experience (which is supported by contentions reported by Kersta) reinforces my faith in the hypothesis that individuality will out — even in the cases of the most socially acclaimed mimicry. Mimicry provides, and capitalizes on, the "caricature" aspect mentioned in my reply to Dennis Fry, which see.

I do compliment Dr Fourcin for his own highly interesting and pertinent LARYNGOGRAPH work and for his continued cautiousness in Voice Recognition regards.