ON A CLASSIFICATION SCHEME FOR SPEECH RATING STUDIES

by Wil P.F. Fagel

1. INTRODUCTION

At the Institute for Phonetic Sciences of the University of Amsterdam rather extensive research has been done on the judgment criteria for the pronunciation of Dutch (Blom & Koopmans-van Beinum, 1973; Blom & Van Herpt, 1976). This has led to a follow-up investigation to construct and validate a standardized procedure for rating the quality of voice and pronunciation and, eventually, to determine the relation of the obtained judgment criteria with physical (phonetic) parameters measured in samples of speech.*

For this investigation a host of literature was collected referring to speech ratings. To determine the relation of each reported study to our, so-called, ONU investigation, we propose a classification scheme for these studies, which we outline below.

2. SPEECH RATING STUDIES

Generally speaking, all those experiments whose 'input' consists of the presentation of speech and whose 'output' is composed of listeners' ratings provoked by these speech samples, can be interesting for the investigation of judgments on the voice and pronunciation of the speaker. We refer to experiments

*This research is supported by the Netherlands Organization for the Advancement of Pure Research (ZWO), project nr. 17-21-13.

satisfying these general conditions as SPRAT (speech rating) experiments.
Studies not demanding a certain aspect of speech itself to be rated, are explicitly included here, since judgments of the speaker's personality or of the speech situation, for example, are still based on speech samples. Such studies can help us to find relevant sociolinguistic and paralinguistic scales, when needed.

3. THE INPUT

3.1. Grossly, the input of SPRAT experiments can be rated differently due to at least five discernible causes (effects of interaction here are left out of consideration):
1) intra-individual or within speaker (WS) differences.
2) inter-individual or between speaker (BS) differences.
3) manipulation of the speech signal or channel (CH) differences.
4) intra-individual or within listener (WL) differences.
5) inter-individual or between listener (BL) differences.
Each of these types of differences can occur on a phonetic (PHON), syntactic (SYN) or semantic (SEM) level. However, on the SYN and SEM levels it is more adequate to speak of 'message' instead of 'channel' differences.

3.2. The ONU investigation is aimed specifically at the variability in judgments caused by differences among speakers on the phonetic level (BS/PHON variability). In assessing this variability, the variation caused by the other input variables is only disturbing and must be considered 'error variation'. Notice that here we use the word 'variation' and not the more technical 'variance'; this was done intentionally, because the latter term usually presupposes the total output variation to be additively composed of those 'variations' which are caused by the different input variables per se; other rules

of combination are possible however. To put it more simply: variance is typically additive, and that is open to question regarding the variance we might encounter.

3.3. It might be worth-while also to consider those SPRAT studies concerning the variability at the phonetic level of which the object is not primarily the BS/PHON variation. Except their possible methodological relevance they might hand us interesting ideas about dimensions used for rating voice and, consequently, yield useful scale terms. Besides, those studies might contain clues how to control the -for ONU disturbing- sources of variation.

3.4. Of course this last remark is also valid for SPRAT studies on the SYN and SEM levels, but the disturbing variation at these levels can -assuming it to be significant- be controlled easily by letting all speakers read out the same text (we then might get involved with variation between speakers concerning ability to read aloud, but that is another question). Consequently, syntactic and semantic SPRAT studies hardly seem more relevant for ONU than perception studies in general are, that is to say, purely methodological.

3.5. Limiting ourselves to SPRAT studies on the phonetic level, we can classify these studies as WS/PHON, BS/PHON, CH/PHON, WL/PHON and BL/PHON, according to the character of the independent variables involved.

3.6. In WS/PHON experiments the variability in the judgments of the speech of a single individual has its origins in the speaker himself (ideally). He (or she) can utter a sentence, for example, in different ways according to rules which either form relatively well-integrated parts of the language system (stress, intonation, rhythm), or can be considered paralinguistic rules (breaking voice, for instance); such variations give

evidence of the speaker's emotions and intentions at different moments.

One single speaker can also act as if various voice and pronunciation characteristics are stable, such as hoarseness, a high-pitched voice or different accents and dialects.

Investigators of the perception of accent and dialect usually try to eliminate the effects of the more idiosyncratic speech characteristics like speech rate, loudness, timbre and pitch. This is often achieved by the so-called matched-guise technique (Giles & Bourhis, 1976): different accents, dialects or even languages are produced by one single speaker, but presented to the listeners as belonging to different persons. So the WS effect here is looked at as a BS effect, in which variation the individual voice characteristics are neutralized. Henceforth we will treat such simulation experiments as BS/PHON experiments.

Should the WS/PHON variation be considered error variation, then it can be controlled "by the choice of speech materials used, by instructions to the speaker, and, more generally, by the circumstances of the recording situation." (Voiers, 1976, p.4).

3.7. In CH/PHON experiments the speech signal is distorted, for instance by chopping off certain frequencies or by channel-inherent noise. These experiments are typically conducted in testing the usefulness of various systems of communication. Sometimes the speech signal is manipulated by varying systematically its rate, mean fundamental frequency or variance of fundamental frequency, but also by interchanging certain voice parameters (hybrid voices; cf. Matsumoto et al., 1973) or by replacing parts of the natural speech by synthesized speech (certain vowels, for example). It is characteristic for these studies that ratings are done on "intellegibility", "accepta-

bility", "degradation" or some other term referring to "communication quality" (Voiers, 1976) and that it is not so much the speech but rather the channel being rated, like in telephone research and in research of hearing aids (cf. Witter & Goldstein, 1971). Strictly speaking, ratings of esophageal and artificial larynx speech (cf. Bennet & Weinberg, 1973) must be reckoned to the BS studies, however 'deformed' this speech might sound: there is no question of an actual distortion, though there is of a deviant mode of speech production.
In some experiments systematically deformed speech is presented to the listeners as coming from different speakers (cf. Brown et al., 1974). We can treat such experiments as special cases of BS/PHON research for the same reason as pertained to the matched-guise experiments (see § 3.6).
Again, other CH/PHON experiments can be considered 'disguised' WS/PHON studies. Uldall (1960), for instance, synthetically imposed differing intonation contours on the same sentence, spoken by one speaker. Her subjects rated each sentence-plus-intonation as to whether it conveyed the impression that the speaker was bored or interested, rude or polite, agreeable or disagreeable, and so on.
In case the CH/PHON effect only amounts to disturbing variation, its occurence should be prevented as much as possible by only using speech recordings of equivalent sound quality, which for all raters should be played back with the same equipment in optimally equal acoustic rooms. With present-day's technology these requirements should hardly offer us any serious problems.

3.8. In WL/PHON experiments the same speech sample can be judged differently by one listener at various moments, not only on account of a 'norm effect', caused by the usual fluctuations in the rating process (discriminal dispersions), but also dependent on such factors as fatigue, health and

effects of learning (adaptation).
WS/PHON experiments can be the topic, for example, of psycho-acoustic studies on discrimination learning and usually bear on very small pieces of the speech signal (so-called phonemes). The norm effects, if disturbing, should be controlled by a statistically warranted replication. Effects like that of learning can be partly eliminated by first offering the subjects some trial stimuli to get adapted and by a balanced presentation of the stimuli with regard to their sequence and number.
Voiers (1976) larded his experimental stimuli with two fixed reference stimuli (one obviously scoring low, the other high on acceptability) to measure the adaptation effect in acceptability ratings of system processed speech and to adjust the responses of his subjects accordingly. Each stimulus was offered several times, in all possible contexts. This implied that every experimental stimulus was once preceded and once followed by every reference stimulus. Voiers noticed, however, that in the long run raters seemed to recognize the anchors and consequently rated them very consistently, independent of the other acceptability ratings.

3.9. In BL/PHON experiments the way listeners differ in their speech rating is examined and to which degree they do so. The influence of BL differences particularly is involved when the judgments are, partly at least, a matter of personal taste or preference. It should be kept in mind, however, that the same differences pertaining to one listener at various moments can also appear within different speakers simultaneously.
Some listener qualities influencing speech perception are: hearing condition, previous training in listening, language background, set, and motivation. Because of selection and long-term effects of learning, audiologists, for instance,

will probably rate speech differently from 'naive' listeners. If BL/PHON effects are considered undesirable, then the most direct method to suppress this 'error' is to extend the number of listeners. Voiers (1976) however, points out another possibility: individual differences in response tendency can be rated independently to supply a statistical basis for the adjustment of data obtained from 'deviant' judges.
Listener ratings of a set of standard conditions, for example, can be used to determine the extent to which a listener shows a tendency to judge more mildly or more severely than the typical or normative listener. His or her responses on the experimental conditions can then be adjusted accordingly.
Voiers applied to this method as a consequence of the fact that he wished to develop a standardized test to examine the quality of diverse systems of communication with only a few raters (which is faster and cheaper).
It is very well possible, for that matter, that one group of raters attends to quite different aspects of speech (judges along different dimensions) then another group, either because one group perceives speech in a very specific and pronounced way (usually due to profession; speech therapists vs. naive listeners*, for example), or because the members of one group share more general (social, regional) features which might influence their speech ratings, cf. Giles (1971), who found a tendency towards "accent loyalty" (judges perceived voices representative of their own speech community in particular respects more favourably than the other regional accent presented). Unless one is only interested in the rating behaviour of the examined group, these BL group differences should be thoroughly reckoned with.

3.10. We now have arrived at the BS/PHON experiments to which ONU belongs. With these studies it is attempted to examine

---

*Bock et. al. (1977) call this phenomenon "fallacy of expert opinion".

speech rating differences (on the phonetic level) purely as a function of speaker differences. These differences arise from diverse aspects of the soundmaking process -like articulation and phonation- and can be described as differences in voice, pronunciation, accent, dialect and the like.
Before we continue to distinguish between different BS/PHON experiments, we will have a short look at the possible ways of rating.

4. THE OUTPUT

The output of SPRAT experiments consists of all possible ratings elicited by speech samples presented to the listener. These judgments may bear directly on speech on the phonetic (voice/pronunciation), syntactic (structure) or semantic (content) level. But also, by way of the perception and evaluation of speech (based on one of the 3 levels, or more globally), ratings may be done of:

- either the relative stable and lasting features of the speaker, such as personality characteristics, attitudes, social status and region of origin; summarizing we call these judgments: ratings of the person.

- or the more temporal and transient features of the context in which the speech signal is being produced; hereby we imply judgments about the speaking situation, the emotions and intentions of the speaker etc., summarized under the heading: ratings of the situation.

For ONU the most interesting studies in literature are those, which supply judgments based on phonetic features. Now we can distinguish between such judgments those that are directly (D),

those concerning the person (P), and those concerning the situation (S). Grossly, we can state that P-ratings bear on BS differences, and S-ratings on WS-differences. P-ratings and S-ratings can -as mentioned before- be interesting with regard to the supply of sociolinguistic and paralinguistic scales. D-ratings can bear on BS differences as well as on WS differences; the same rating terms can be used here as for the other ratings (P & S), if only the listeners are explicitly instructed to rate the speech itself. Indeed, from this it becomes clear that the above distinction between different types of rating is rather arbitrary; however, this distinction and that between the different types of input variations serves to classify SPRAT experiments in literature more quickly, by which means their relevance for ONU can also be assessed more efficiently.
D-ratings can vary from perceptual similarity ratings of separate speech sounds regarding some specified feature up to global evaluative judgments of lengthy speech fragments.

5. ONU AND THE BS/PHON EXPERIMENTS

ONU belongs to the BS/PHON studies using global D-ratings. Looking for other BS/PHON experiments in literature, we often find them to be difficult to compare with the ONU investigation, because either loose sounds (sustained or not, sometimes even sung) are presented instead of a spoken text, or small ('gated') pieces of the speech signal. Also, these experiments often contain presentations of pathological speech, or strongly deviating speech (judgments on articulatory proficiency in second language learning, for instance). Finally, the speech signal sometimes has been presented in combination with other clues (by use of videotape, for exam-

ple). This fact prohibits us to look upon the experiment as a pure SPRAT study.
At the input side we can now formulate the following restrictive conditions which should be met by any SPRAT experiment to be sufficiently comparable with ONU.
The presented speech should be:

a) a spoken text and no sustained sounds nor sung texts;

b) not 'gated', i.e. no small pieces, but longer fragments of speech;

c) purely auditory, i.e. independent of other, for instance visual, clues and of previous knowledge with regard to features of the speaker or the speaking situation;

d) not pathological (no cleft-palate speech, no pathological hoarseness etc.) and not strongly deviating (for example, speech produced by someone whose mother tongue obviously is not the present language).

On the other hand, just like BS/PHON experiments with P-ratings and S-ratings, studies that violate one or more of the above restrictions can still be worth looking at with regard to the supply of (suggestions for) rating terms. In most cases, however, it turns out that very experiment-specific rating scales have been used, which hardly have any relevance for ONU.

6. CONCLUDING REMARKS

As stated before, the distinctions made between different methods of speech presentation, along with the various purposes these might serve, and between the different judgments

that can be given consequently, offer us a frame to classify the host of reports on speech rating experiments. This considerably facilitated our search for literature relevant to the ONU investigation (an account of the actual survey of this literature will be given elsewhere) and we hope other researchers in the speech rating area can also profit by the proposed classification scheme outlined above.

REFERENCES

Bennet, S. & Weinberg, B. (1973). Acceptability ratings of normal, esophageal, and artificial larynx speech, Journal of Speech and Hearing Research 16, 608-615.

Blom, J.G. & van Herpt, L.W.A. (1976). The evaluation of jury judgments on pronunciation quality. Proceedings 4, IFA, 31-47.

Blom, J.G. & Koopmans-van Beinum, F.J. (1973). An investigation concerning the judgment criteria for the pronunciation of Dutch, Proceedings 3, IFA, 1-24.

Bock, G.B., Powell, L. & Flavin, J.W. (1977). The influences of sex differences in speech evaluation: situational and media effects, Communication Education 26, 143-153.

Brown, G.B., Strong, W.J. & Rencher, A.C. (1974). Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech, J. Acoust. Soc. America 55, 313-318.

Giles, H. (1971). Patterns of evaluation to R.P., South Welsh and Somerset accented speech, British Journal of Social and Clinical Psychology 10, 280-281.

Giles, H. & Bourhis, R.Y. (1976). Methodological issues in dialect perception: some social psychological perspective. Anthropological Linguistics 18, 294-304.

Matsumoto, H., Hiki, S., Sone, T. & Nimura, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates, IEEE Transactions on Audio and Electroacoustics 21, 428-436.

Uldall, E. (1960). Attitudinal meanings conveyed by intonational contours, Language and Speech 3, 223-234.

Voiers, W.D. (1976). Methods of predicting user acceptance of voice communication systems, Final Report, Contr. No.DCA100-74-C-0056, D-76-001-U.

Witter, H.L. & Goldstein, D.P. (1971). Quality judgments of hearing aid transduced speech. Journal of Speech and Hearing Research 14, 312-322.