SUMMARIES OF PH.D. THESES DEFENDED IN 1992

ON MEASURING AND PREDICTING SPEECH INTELLIGIBILITY

author: Herman J.M. Steeneken[1]
promotor: L.C.W. Pols
co-promotor: T. Houtgast[2]
date of defence: 11 June 1992

Summary

The intelligibility of speech, degraded by a speech-communication system, has been the topic of many studies in the past 70 years. Already between 1920 and 1930, Fletcher and Steinberg developed several methods to determine intelligibility. They found a relation between the transmission quality and several physical aspects of the transmission channel. Mainly bandwidth and signal-to-noise ratio were considered. The second world war had a great impact on the evaluation of speech communication. Many papers appeared just after the war on the subjective and objective assessment of speech-communication systems (Egan, 1944; French & Steinberg, 1947; Beranek, 1947; Fletcher & Galt, 1950).

Subjective intelligibility measures

The intelligibility of sentences is an obvious measure for quantifying the quality of speech communication. However, a sentence-intelligibility score already reaches 100% at a poor-to-fair transmission quality and is therefore limited in its use. A more generally applicable measuring method is based on nonsense syllables of the CVC-word type (consonant-vowel-consonant). This type of test discriminates between transmission conditions over the full range from bad to excellent and was also used in the present study (chapter 3).

The significance of the test results for several parameters of the test conditions such as speakers, listeners, and speaker sex was analyzed. The relation between the CVC-word scores and the individual phoneme-group scores (initial consonants, vowels, and final consonants), and phoneme types (fricatives, plosives, vowel-like consonants, and vowels) was also analyzed. To support the grouping of certain phonemes, a principal-component analysis on the phoneme scores for a wide variety of conditions, as well as multi-dimensional scaling on the phoneme-confusion matrices, were used. It was found that at the phoneme level, four groups can be identified, each with a fairly similar distribution of responses for various transmission conditions. The differentiation between the phoneme-group responses can be used as a diagnostic tool and to improve predictive intelligibility measurements.

---

1 TNO Institute for Perception, Soesterberg, The Netherlands
2 idem

We obtained, for a number of reference conditions, the optimal relation between sentence intelligibility and phoneme-group scores. A reliable prediction of sentence intelligibility was obtained by a weighted combination of phoneme-group scores.

Objective intelligibility measures

Rather than measuring intelligibility of degraded speech with subjective intelligibility measures, the effect on intelligibility of a transmission channel can also be predicted by considering the physical properties of such a channel. The accuracy of this prediction as obtained by two existing models (AI and STI) could be improved.

The Articulation Index (AI; French and Steinberg, 1947) and Speech Transmission Index (STI; Steeneken and Houtgast, 1980) are based on a linear summation of the contributions of individual frequency bands to intelligibility. There is evidence that this assumption that frequency bands are independent from each other is not correct for conditions with gaps or selective masking in the frequency domain.

We designed an experiment in which the contribution of individual frequency bands, and the question of mutual dependency, could be studied. Evaluation of the observed scores and the corresponding physical specifications resulted in a *revised model*, which accounts for the mutual dependency between adjacent octave bands by the introduction of a so-called redundancy correction factor. The weighting factors proved to be identical for male and female speech and for various signal-to-noise ratios. This robust model for prediction of intelligibility gives a significant improvement of the prediction accuracy in comparison to the original model (chapter 2).

However, the optimal frequency weighting and redundancy correction also depends on the type of speech considered, Therefore, the four groups of phonemes with a similar response at various transmission conditions (fricatives, plosives, vowel-like consonants, and vowels were used to specify the frequency weighting for various types of speech. For each group the optimal set of frequency-weighting factors and the optimal redundancy-correction factor were determined separately (chapter 4).

The predicted phoneme-group scores can be used to predict the word score of several types of nonsense words. This is performed in two steps. For instance, to calculate the CVC-word score, the scores of the initial consonant, the vowel, and the final consonant are calculated first by a summation of the phoneme-group scores (weighted according to the frequency of occurrence of the phonemes). The word score is obtained by the product of the (normalized) consonant and vowel scores.

The revised STI_r model for the prediction of intelligibility was validated for a set of independent transmission channels. It was concluded that the present STI_r model provides a reliable measure, applicable to a wide range of transmission conditions (chapter 5).

HMM-BASED CONTINUOUS-SPEECH RECOGNITION. SYSTEMATIC EVALUATION OF VARIOUS SYSTEM COMPONENTS

author: Paul van Alphen[3]
promotor: L.C.W. Pols
date of defence: 3 July 1992

Summary

In this thesis a description is given of our automatic speech recognition system (called REXY). This system is capable of recognising continuously-spoken sentences, and transcribing these sentences into written text. With the REXY system experiments have been performed to evaluate various system components.

Speech recognition by computer is only one aspect of voice interaction between man and computer. The counterpart of speech *recognition* is speech *synthesis*: with speech recognition the computer is *listening* (speech input), while with speech synthesis the computer is *speaking* (speech output). In a dialogue situation between man and computer, the computer should be able to perform both speech recognition and speech synthesis. Another "listening" task for computers, that is often mistaken for *speech* recognition, is *speaker* recognition: with speech recognition the computer tries to determine *what* is spoken (words or message) and with *speaker* recognition the computer tries to find out *who* is speaking. Within the speech recognition domain we restrict ourselves to speech-to-text transcription. This means that we start from the speech sound and then try to find (recognise) the words that have been spoken. Extracting the meaning of an utterance requires much more linguistic knowledge and interpretation.

In a speech recognition system we distinguish the following two stages: acoustic preprocessing and classification. Acoustic preprocessing starts with the digitalisation of the speech sound and this digitised signal is transformed into so called feature vectors. The classification algorithms takes the feature vectors as input, and tries to find (recognise) the words that have been spoken.

The REXY system is able to recognise continuously spoken (Dutch) sentences. Since the system is trained with phones, any Dutch word can easily be incorporated in the system, although in the training sentences only 238 different words occurred. Furthermore, for practical reasons the present system is trained and tested as a speaker-dependent recogniser for one male speaker only. In Chapter 6 (and Appendix D) the database with speech utterances is described. With the REXY system we systematically evaluated various system components. Details about the experimental procedures can also be found in Chapter 6. The conclusions that can drawn from these experiments are elaborated in Chapter 7 and are summarised below.

One of the components we varied was the acoustic preprocessing. We have investigated two types of analysis, and we have experimented with several feature vectors (a description of the preprocessing is given in Chapter 2).

---

3 Presently at PTT Research Laboratories, Leidschendam, The Netherlands

* The two types of analysis are a filterbank and a LPC analysis. The overall performance of the filterbank analysis turns out to be the better of the two.

* Our experiments showed that the performance of the recognition system could benefit from the cooperation of different feature vectors. The best performing combination is the filterbank preprocessing together with three feature vectors: the "slope" vector (frequency derivative of the filterbank spectrum), the time derivative of the slope, and the time derivative of the energy (see Chapter 2 for details).

The classification algorithm we used is based on discrete HMM (hidden Markov modelling) technology. In three subsequent chapters this classification algorithm is described, as well as a dynamic programming technique used to perform an integrated search. Markov theory (Markov chains and Markov processes) is introduced in Chapter 3. In Chapter 4 we expand the Markov models to hidden Markov models and in Chapter 5 we adapt the hidden Markov models to model speech. As unit of modelling we chose for the Dutch phonemes. By applying the dynamic programming, the Markov models are combined with a word-duration model and a grammar model. The experimental results allowed us to draw the following conclusions:

* Initialisation of the HMM parameters must be done with care. We compared a uniform (all parameters have initially the "same" value) and a sophisticated way of initialisation (based on hand-segmented data). Sophisticated initialisation yields a system that has a better recognition performance.

* As long as the HMM parameters are not well trained (which is almost always the case in actual conditions and which was also the case in our experiments), smoothing of the parameters is important. The smoothing technique we implemented is called "cooccurrence smoothing" (this technique smoothes the probability density functions of the Markov models).

* Because the HMM's do not model duration very well (only implicitly), we tried to model the word duration explicitly with a Gaussian distribution. The recognition benefit of this kind of duration modelling turned out to be limited.

* Dynamic programming integrates knowledge about the spoken words (in the HMM's) with a simple grammar model. Different "bigram" grammars have been implemented with "perplexities" 60, 20, and 2.4 (lower perplexity implies a stricter grammar). The effect of the grammars is large: the error rate reduced from 25.8% (for the "no grammar" case with a perplexity of 110) to 15.5% (perplexity is 60), 5.3% (perplexity is 20), and 0.9% (perplexity is 2.4) given filterbank preprocessing.

* The grammar model and the word-duration models can simply be integrated with the Markov models (Viterbi search). This means that at recognition time an integrated search is performed with many knowledge sources: acoustic and phonetic knowledge from the HMM's, lexical knowledge from the (word) pronunciation dictionary, word duration, and syntactical knowledge from the grammar.

The experiments we performed with the REXY system indicate that high recognition performance can only be achieved if preprocessing and classification are both performed adequately. In designing a recognition system, both preprocessing and classification have to be optimised and tuned to each other.

MALE AND FEMALE SPEECH. AN EXPERIMENTAL STUDY OF SEX-RELATED VOICE AND PRONUNCIATION CHARACTERISTICS

author: Mirjam T.J. Tielen
promotor: L.C.W. Pols
co-promotor: F.J. Koopmans-van Beinum
date of defence: 17 September 1992

Summary

There are numerous indications that people extract more information from speech than simply the message itself. We are able to identify speakers by their voice and pronunciation, to recognize their regional background, their mood, and several other characteristics.

Generally, we can also identify the sex of the speaker from his/her voice and/or pronunciation. Women speak with a relatively high-pitched voice and men with a low-pitched voice. The differences regarding pitch height are related to differences between the sexes in the anatomy and physiology of the vocal apparatus. However, apart from pitch height, little is known about phonetically-related differences between men and women.

The reason why some people speak more quickly, more melodiously, more broadly, or with more authority than others seems to be determined by environmental factors rather than by biological factors. People tend to adapt to their role in society regarding their clothing, their way of acting, and also their way of speaking.

It is common knowledge that men and women play, or at least are more or less expected to play, different roles in our society. E.g. children-caring is done especially by women, while jobs with management aspects are taken most frequently by men. Such expectations or norms towards men and women may also influence the speech production and speech perception behaviour of men and women.

The distinction between speech of men and women is also apparent if one considers the developments in speech technology. In speech synthesis as well as automatic speech recognition there is a clear preference to use 'male-like' voices, whereas it is not clear at all, except for a few characteristics such as pitch, to what extent the voice and pronunciation characteristics of men and women differ.

The main aim in the present study was extracted from the above mentioned arguments. The aim was to obtain more insight into the voice and pronunciation characteristics of men and women, while distinguishing between attributed and actual characteristics of men and women (ch. 1). The attributed characteristics were measured by means of introspective judgments, whereas the actual characteristics were measured by means of perceptual or acoustic analyses.

Three main topics were chosen with respect to possible differences between speech of men and women. The first topic was the evaluation of voice and pronunciation characteristics by means of semantic scales. The second topic was pitch/fundamental frequency and the third topic was the intelligibility on the level of words and phonemes.

The description of our study is started with two experiments in which the importance of non-verbal cues in speech was tested (ch. 2). Firstly, an identification experiment is described in which the ability of listeners to extract information about age and sex from

voice and pronunciation cues alone was examined. It appeared that the listeners were very well able to identify the sex of the speaker, but also to classify the age (which is less obvious).

Secondly, an introspective experiment is described in which judges gave their opinion about ideal and average voice and pronunciation characteristics of men and women, by means of semantic scales (without actual presentation of speech). Regarding the characteristics of *ideal* voice and pronunciation, it was found that the differences between men and women were restricted to the fact that the ideal female voice should be higher and softer than the ideal male voice. Regarding the characteristics of *average* voice and pronunciation, the judges indicated far more differences between men and women. Also, it was found that the expected *average* characteristics for male speakers appeared to be closer to their *ideal* characteristics than those for female speakers.

Introspective judgments reveal insight into the norms and expectations with respect to voice and pronunciation of men and women. However, it could very well be that those ideas are based on sex-related stereotypes and not necessarily due to actual speech performance. Therefore, a listening experiment was carried out in which 40 listeners evaluated voice and pronunciation of 30 men and 30 women, again by means of semantic scales (ch. 3).

Apart from the variables 'sex of speaker' and 'sex of listener', a third variable was included in order to analyse the influence of another factor, which is specifically socio-culturally determined, on voice and pronunciation, i.e. 'profession of speaker'. The speakers were representatives of one out of the following profession categories: nurses, managers, and information agents (with equal numbers of male and female speakers in this experiment). These professions differ with respect to socio-economic status (SES) as well as with respect to the actual distribution of men and women over the three professions.

A number of characteristics appeared to differentiate between male and female speakers. However, these distinctions were not always in agreement with the literature or with the introspective judgments mentioned above. In the literature it is e.g. suggested that women speak in a more polished way than men and men speak with more authority than women. In contrast to this, our perceptual data reveal that male and female speakers sounded equally polished and authoritative. The data further indicate that the professions were clearly differentiated from one another with respect to characteristics of voice and pronunciation. Moreover, the significant differences are in agreement with stereotypes of these professions (e.g. managers speaking in a distinguished way and nurses speaking sweetly).

From the foregoing it is clear that the listeners had differentiated between the sexes and the professions without any other clues than voice and pronunciation. Subsequently, an identification experiment was carried out in order to examine whether or not listeners are able to classify the professions correctly. The results show that this is indeed possible.

Apart from perceptual evaluation, also introspective evaluation was executed about voice and pronunciation characteristics in the three profession categories, separately for men and women. Those results show for instance that women were supposed to speak in a more polished way than men, whereas this tendency was not at all present in the perceptual evaluation. Regarding the different professions, it appears that only partly the same tendencies are found as for the perceptual evaluations.

In addition to the perceptual and introspective evaluation by a large group of judges, also the opinion of the 60 speakers themselves about their own voice and pronunciation was asked. The results of that evaluation show no significant differences, neither between male and female speakers nor between the professions. So, the speakers

themselves seem not to be aware of their distinctive voice and pronunciation characteristics.

The second topic was pitch/fundamental frequency (ch. 4). (We use the term 'pitch' when considering the perceptual domain; the term 'fundamental frequency' is used when referring to the acoustic domain).

In the literature, as well as by our listeners and judges, it was reported that pitch is the most salient factor for distinguishing between speech of men and of women. However, is this restricted to mean pitch/fundamental frequency or do the range and variation of pitch/fundamental frequency also play a role? From the above mentioned evaluation experiments, the general tendency in this respect was that female speakers sounded more melodious than male speakers. This might imply that more fundamental frequency variation is present in speech of women.

Acoustic analyses were carried out for several read speech samples of groups of male and female speakers. As was expected, the data reveal a clear difference in *mean* fundamental frequency between male and female speakers (±120 Hz versus ±200 Hz, respectively). No significant differences in mean fundamental frequency were found between speakers with a different educational level or different profession. It is striking that the different speech conditions under study (sentences and text) also resulted in similar mean fundamental frequency values.

Although considerable differences were found between the individual speakers with respect to fundamental frequency *range* or *variability*, no differences were found between the two sexes. Also, with respect to the factors 'educational level' or 'profession' no differences were found in fundamental frequency range or variability. Of course, our results are to be restricted to the reading condition.

The relationship between acoustics and perception is rather clear as far as pitch height is concerned. However, only low correlations were found between fundamental frequency range and variability on the one hand and judgments regarding melodiousness and expressiveness on the other hand. Did we catch the wrong acoustic parameters for obtaining useful information about pitch variation (intonational) aspects? In order to verify the difference in fundamental frequency patterns between men and women, a perception experiment was carried out in which manipulated speech was presented to listeners.

The results indicated that the subjects had not been able to identify the sex of the speakers by means of information about fundamental frequency range and variability alone. So, the conclusion must be that at sentence level, fundamental frequency variability plays a minor role for sex identification.

With regard to the third topic, i.e. the effect of speaker sex on intelligibility, contrasting suggestions have been found. For instance, a strong preference exists for male voices in speech technology applications, while on the other hand there is a preference for female voices in actual announcement situations (e.g. in department stores).

Intelligibility was measured in several noise conditions (ch. 5). Ten male and ten female speakers of Standard Dutch were selected. In terms of Consonant-Vowel-Consonant (CVC) words, it appears that the group results for male and female speakers show equal word and phoneme intelligibility under all noise conditions. The differences between the individual speakers were rather large. Evaluation of the intelligibility of all speakers by means of the semantic scale 'low intelligibility - high intelligibility' revealed similar results with respect to the rank order of the different speakers.

The phoneme confusions were also analysed. However, no fundamentally different patterns were found for male as opposed to female speaker data. Most confusions took place between phonemes that differed only with respect to one distinctive feature.

The aforementioned results do not indicate any striking difference between men and women with respect to voice and/or pronunciation. In general, it can be concluded from our study that less actual (perceptual or acoustic) differences with respect to voice and pronunciation characteristics of men and women were found than were indicated in the literature or attributed by judges (ch. 6).

Regarding the socio-culturally determined characteristics, the differences between male and female voices and pronunciation which were actually (perceptually or acoustically) found, seem to be of the same order as the differences found between the professions under study. In that case, the distinction of speakers between males and females is only one out of several other possible distinctions.

The restriction in our study to the use of read speech meant a clear abstraction from real-life speech situations. We chose for this abstraction in order not to be drowned by uncontrollable variables. However, we hope that future studies in the field of male and female speech will proceed more and more towards natural speech situations.

THE EVALUATION OF SPEAKING ABILITY IN COMMUNICATIVE SITUATIONS: global rating and detailed analysis of oral performance of students of 11 to 12 years of age

author: Amos van Gelderen[4]
promotor: L.C.W. Pols
co-promotor: G.C.W. Rijlaarsdam[5]
date of defence: 16 November 1992

Summary

Introduction

What are the main dimensions according to which can and should evaluate the speaking ability at the end of primary education? This is the question that guided the studies reported here. The question arises in the context of a National Assessment of Educational Performance. This survey aims at several goals. First, it is intended to inform the public about the effectivity of language education. Second, it wants to provide an empirical basis for the discussion about educational level and whether it needs to be improved. Third, it is directed to provide educators and educational researchers with means for educational improvement. In order to fulfil these goals satisfactorily, the testing devices that are used must provide a rich source of information. It will, for example, not be sufficient to inform the public that the speaking ability of students in Holland is 'poor'. In other words, we need more precise information about which aspects of the oral performance are disappointing, under which conditions the results are obtained and how they can be related to educational improvement.

On the other hand large scale surveys improve restrictions on the administration of tests, especially tests for oral performance that are individually administered. Moreover, rating procedures require the use of trained assessors, which is rather costly and time consuming. Therefore I undertook to develop and test a general rating scheme for the evaluation of speaking ability that results in reliable and valid ratings of different aspects of the ability and at the same time satisfies the requirements of efficiency in large scale assessments. A central assumption in the assessment of speaking ability in the context of primary education is that the most appropriate condition for testing is the simulation of realistic communication. That is, the testing situation, the so-called integrated task, should consist of a communicative purpose against the background of a real-life situation that students recognize as such. Accordingly, **criteria** for assessment should derive the communicative effectiveness of speech. These assumptions are based on the fact that language education primarily aims at providing the necessary skills to participate in all kinds of communicative situations. So the task of a national assessment is to evaluate to what extent the educational system succeeds. This poses specific problems for a valid evaluation. Which types of communicative tasks are relevant for the assessment of speaking ability of students of a certain age? How many different tasks should be administered and how varied will they have to be to provide a

4 S.C.O.- Kohnstamm Institute Amsterdam, The Netherlands
5 Instituut voor de lerarenopleiding Amsterdam, The Netherlands

satisfactory coverage of the domain? Although the main purpose of the empirical studies reported was to develop and validate a rating scheme, the so-called problem of task validity could not be ignored. It soon appeared that evaluation criteria are to some extent dependent upon the characteristics of (integrated) tasks. Moreover the applicability of the rating scheme had to be limited from the beginning: only in tasks where individual speakers - instead of pairs or trios - can be rated for their contribution to the communication, use of the scheme will be warranted.

Data collection

Students of the last year of primary education performed on four oral tasks: two tasks were narrative, one task consisted of alarming the police by telephone and one task of an exposition of the way a spider builds his web. Except for the alarming task, in all tasks classmates functioned as listeners.

Sound recordings were made of all performances. Data collection took place in two different samples. One sample consisted of two hundred students and can be regarded as a nationally representative sample; the second sample consisted of one hundred students from the region of Amsterdam and surroundings. The registration of the oral performances in general was of an acceptable quality for assessment purposes. In view of the intended validity study and the phonetic analyses that had to be carried out, special care was taken in the recording sessions for the second sample.

Theoretical foundation

A rating scheme is proposed consisting of four functional dimensions. These are based on an overview of so-called analytic schemes that have been developed in studies of the rating of speaking skills in diverging contexts (Wesdorp, 1981). These dimensions are defined by functions that can be derived from the general criterion of communicative efficacy. Two dimensions - **Reference** and **Delivery** - are directly related to communicative content. Reference is defined by the representational function of language; Delivery is defined by the functions of expression and appeal (Bühler, 1982). The dimensions interchangeably - dependent upon the communicative situation - denote the dominant communicative functions that are to be realised. The other two dimensions - **Fluency** and **Intelligibility** - are indirectly related to communicative content and apply to the conditions that have to be met in order to produce interpretable utterances. Fluency is defined by the realisation of continuity of speech and Intelligibility by the quality of the realisation of utterances ('decodability') (Crystal & Davy, 1979).

In order to use the four dimensions as a rating scheme, each dimension is regarded as a heuristic device from which specific criteria for assessment in a given speaking situation can be deduced. Furthermore a linkage is assumed between the criteria deduced from the dimensions on the one hand and the aspects of behaviour that are the objects of assessment on the other. Specifically, for Reference only linguistic aspects are seen to be relevant, for Delivery linguistic, phonetic and non-verbal aspects are relevant, for Fluency and Intelligibility linguistic and phonetic aspects. On a more concrete level, however, it is supposed that the same aspects of behaviour **do not** always serve the same functions. That is why the differentiation of the dimensions is solely based upon the communicative **functions** to be evaluated and not upon the precise behavioural **aspects** that can be distinghuised.

Empirical test of the rating scheme

The rating scheme has been put to empirical test in two steps. First, several rating categories have been derived from each dimension and have been applied in small scale experiments by jury's of four or five raters. In these experiments (N=40) performances of students on the four tasks, selected from the larger data set, are rated after an

instruction- and training-session. The purpose of these experiments is to acquire knowledge as to the applicability of the rating categories for performances on different oral tasks, the degree of consensus among raters, the instrumental differentiation that exists between jury-ratings of different categories and optimal rating conditions (rating several categories simultaneously vs each category separately). Second, on the basis of these experiments, a more definite test of the scheme has been carried out. A jury of three raters (all women with experience as teachers in primary education) applied selected categories - one for each dimension of the scheme - to rate the performances of all students in our two samples on the four oral tasks. In both steps - the small scale experiments and the large scale studies - categories have been derived from the dimensions in a taskspecific way. That is, categories for the same dimension but applied in a different task often consist of different criteria and require different behavioural aspects to be observed. This is a consequence of a functional - instead of behavioural - rating scheme.

The results of the empirical investigations can be summarized in the following four points.

1. Reliability of the rating categories is at an acceptable level (about .80) for the purposes of a national assessment, when jury's of three trained raters are used.

2. A four-factor model for the correlations among jury ratings, each factor representing one of the dimensions of the rating scheme, fits reasonably well. Furthermore there are strong indications that ratings of categories derived from the same dimension hardly convey distinct information about speaking ability in a given task, whereas ratings of categories derived from different dimensions, although sometimes strongly correlated, do convey distinct information.

3. The rating scheme proves to be applicable for performances on all four tasks tested, but there are indications that in two of the tasks (the alarming task and the exposition) rating of categories for Delivery and Fluency is more difficult, due to short duration of the performances and/or to the lack of cohesiveness of the texts produced.

4. An efficient rating procedure is feasible; hereto a jury of three trained raters rates each performance on four categories simultaneously, provided that the performances are of reasonable length; without significant loss of reliability or validity.

Empirical test of rating validity

A question that was not addressed in the foregoing is whether the rating of the dimensions in oral performances does convey the information about the behavioural aspects stated in the dimensions **definitions**. As mentioned previously, the aspects of speech to be rated are not invariant across tasks. Although from an instrumental point of view it has been demonstrated that ratings of the dimensions convey distinct (but correlated) information, the diagnostic value of these ratings is not yet clear. In short, we cannot exclude the possibility that ratings are based on **other** aspects of the speaking performances than we believe they are, or that the ratings of different dimensions have largely **overlapping** meanings so that their differentiation is invalid. Moreover, some notorious rating problems, such as the 'signific effect' and the 'halo-effect', could have invalidated the resulting scores. To investigate the validity of the jury ratings on the four dimensions, several analyses have been carried out to determine the correlations between these ratings and linguistic and phonetic aspects of the rated performances. In a regression design I tested hypotheses about these correlations. First, these hypotheses state a significant relation between jury ratings and the frequency of the linguistic and phonetic variables that had been mentioned in the definition of the rating dimension in question (the convergent prediction). Second, the hypotheses state that a weaker relation exists between the jury ratings and variables that are mentioned in the definition of **other** rating dimensions (the divergent prediction). The prediction of Delivery and Fluency has received most attention in this

examination, because the differential meaning of those two dimensions has proved to be more problematic than that of Reference or Intelligibility. Therefore a rather large amount of linguistic and phonetic predictors for Delivery and Fluency has been analyzed in comparison with the other dimensions. (Non-verbal variables could not be included because rating of performance occurred from sound tapes). On the other hand, because of the time consuming procedures involved in the analysis of phonetic and linguistic variables, only relatively small selections of performances (sixty per dimension) on one (narrative) task could be analyzed.

For the prediction of ratings of Reference the total amount of relevant 'content elements' has been determined in each performance on three tasks (narrative, alarming and expository) (N=200). Prediction of ratings of Intelligibility has been carried out by calculating the correlation between the ratings and the amount of 'hardly intelligible' utterances in performances on a narrative task (N=60). For the prediction of ratings of Delivery the following variables have been selected: (1) variation of intonation, based on auditory analysis according to a description of fundamental pitch movements in Dutch ('t Hart, Collier & Cohen, 1990), (2) acoustic measurements of variation of fundamental frequency, (3) acoustic measurements of intensity and intensity variation, (4) relative amount of pitch accents (corrected for text length), (5) relative amounts of lexical elements with a positive or negative effect on narrative register. For the prediction of ratings of Fluency the selected predictors are: (1) relative amount of self-corrections and non-functional pauses, (2) duration of self-corrections and non-functional pauses, (3) mean speech rate (pauses included) (4) mean articulation rate (pauses not included). For all variables that can not be measured instrumentally, a detailed coding instruction has been designed and applied by two trained raters. By comparing the codes assigned independently by each rater for the same performances the degree of consensus has been determined. The coding of pitch movements by the two raters has been further examined by comparison with instrumental analyses of a sample of the coded utterances. In all cases coding consensus and accuracy has been found to be satisfactory.

Results show that for the ratings of three dimensions - Reference, Intelligibility and Delivery - the hypotheses can be accepted. The ratings are more strongly related to the linguistic and phonetic variables that are mentioned in their definition than with those mentioned in the definition of other dimensions. The proportion of explained variance of ratings of Reference ranges from 53 to 79 percent (dependent upon the task). Explained variance of ratings of Intelligibility was 37 percent and for Delivery 83 percent. Intonation variation and relative amounts of lexical elements with reinforcing or decreasing effect on register have the greatest part in predicting Delivery. Ratings of Fluency are also substantially predicted (55 percent of the variance of the ratings), however only the duration of self-corrections and non-functional pauses plays a significant role. Moreover it appears that predictors for Delivery also explain a large proportion of the variance of the Fluency ratings (55 percent). Further analysis of the specific meaning of these ratings shows that only rather gross disruptions of continuity of speech are significantly related to Fluency (false starts and pauses of long duration), whereas more subtle hesitations, repeats, filled and unfilled pauses appear to be largely ignored by the jury. Furthermore, no evidence has been found of the occurrence of so-called signific or halo-effects in the ratings of the speech performances. The correlation between ratings of Delivery and Fluency can be largely explained by the correlation that exists between the behavioural aspects rated. Also, no indication has been found for diverging interpretations among raters regarding the relevance of certain behavioural aspects for deciding upon their scores.

Discussion

The results of the empirical studies reported are rather promising. The rating scheme tested proves to satisfy several needs in large scale assessments of speaking ability such as the need to supply differential information about the skills students possess in a reliable and efficient way. Moreover, its utility for the rating of performances on several communicative tasks has been demonstrated. Also, the validity and diagnostic meaning of the rating dimensions was, for the greater part, substantiated. Nonetheless, I must point at some limitations of the studies on which these results are based. First, the sample of students for the validation study for Delivery, Intelligibility and Fluency was rather small, and not nationally representative for the population, so the possibility of statistic generalization is limited. Second, the results are mainly based on the scores given by three trained raters; we can not be certain that other raters' scores are equivalent. Third, several relevant predictor variables for the rating dimensions have not been included in the validity study for various reasons. Fourth, the validity of the rating dimensions has been solely determined on the basis of ratings of performances on one (narrative) task. In view of the dependence of rating criteria and the behavioural aspects to be observed on task characteristics, results can not be generalized to ratings on other types of tasks. Fifth, not all kinds of rating criteria that could be relevant in the assessment of speaking ability in communicative situations have been investigated. Specifically, criteria dealing with standard usage and grammatical correctness or complexity have not been included, although these criteria might be rather important in the case of formal communication. Also, communicative situations in which cooperation among interactants plays an important role, require specific rating criteria that have not been included in our scheme. Criteria for turn-taking and -giving and for evaluation of the process of negotiation and cooperation as such, are important additions if performances on such tasks are to be assessed.

The above limitations all deserve attention in empirical studies. Some of the research themes are specifically important in my opinion. Those themes are elaborated upon. It concerns the following:

1. A redefinition of Fluency on the basis of our validity study. The results of the study have made it clear that the significance of Fluency ratings has been severely narrowed in comparison with the original definition of the dimension; several explanations and implications of this finding are being discussed.
2. The relation between acoustic and perceptive variables in the rating of speech in several empirical studies is discussed. Several occasions in these studies and in the present one are found to speculate about the basis of speech perception and rating: detail or Gestalt.
3. The problem of task validity for the evaluation of speaking ability in communicative situations is explored. What are the main parameters of integrated tasks that have to be varied to reach an acceptable coverage of the domain? A suggestion for an experimental analysis of task parameters is given.

In conclusion, the utility of the rating scheme in two different contexts is discussed: large scale performance surveys and (diagnostic) evaluation in primary and secondary education. A comparison is made with a rating scheme now in use for national performance surveys at the end of primary school and several advantages of the present scheme are pointed out. With respect to in educational contexts it is indicated what advantages there seem to be in using the functional rating scheme in comparison with schoolpractice nowadays. Furthermore, some ideas for implementation of the scheme and some practical consequences for the teachers, the pupils and the curriculum are portrayed.