PHONETIC TRANSCRIPTION AS A MEANS OF DIAGNOSTICALLY EVALUATING SYNTHETIC SPEECH

R. van Bezooijen* and W.H. Vieregge**

* Institute of Phonetic Sciences, Amsterdam, The Netherlands
** Department of Language and Speech, Nijmegen, The Netherlands

ABSTRACT

This paper explores the possibilities of using narrow transcriptions as an enriched alternative to an open response identification test in the evaluation of synthetic speech at the segmental level. To that end, transcriptions of synthesized phonemes were compared with the corresponding identification data. It is concluded that transcription should not be used in place of but rather in combination with an identification test.

1. INTRODUCTION

Probably the best known test for evaluating synthetic speech at the segmental level is the Modified Rhyme Test (MRT) [6], used extensively for the comparative evaluation of American English synthesis systems. In the MRT, initial and final consonants are tested separately with meaningful English CVC words. For each stimulus word the listeners are presented with six alternatives, from which they have to choose the correct response. Although the MRT has several advantages, such as speed and ease of administration to untrained subjects, it has been critisized extensively in the literature, especially with respect to the restrictions imposed on the responses and the limited phonetic contexts in which the target consonants are presented [cf. 3]. The objections raised are particularly serious if the test is to be used for diagnostic purposes, i.e. to assess the flaws of a system with a view of improvement, rather than comparative purposes, i.e. to relate a system's overall performance to that of other systems or other variants of the same system.

An alternative approach, adopted regularly in the diagnostic evaluation of synthesis of European languages [e.g. 2,7] is to use an open response task with a large stimulus set comprehending both meaningful and meaningless words of various structures, such as CVC, VCV, VCCV, and CVVC. In this way, the confusions found reflect true, unbiased perceptual characteristics of the stimulus sounds and information is gained on the intelligibility of phonemes in a wide variety of phonetic contexts. With the right equipment, the responses can be analyzed (semi-)automatically and presented insightfully in terms of percentages correct phoneme identification and phoneme confusion matrices. The subjects need to be trained in the use of an unambiguous notation system, but the time investment can be relatively small if foreign language students are used.

Although the approach described can certainly be considered to be an improvement over the MRT in diagnostic evaluation, one could speculate whether it would not be possible to have an even more finely tuned measuring instrument. For it is not difficult to point out some characteristics of open response identification tests which in their turn limit the type and detailedness of the information yielded. For example, if the subjects perceive more than the intended number of input phonemes, they are forced to make a choice. Also, responses are limited to the phoneme inventory of the language in question. Deviations from standard, natural phoneme realizations (e.g. undue aspiration, excessively abrupt voice onset, inadequate segmental duration) cannot be indicated. Moreover, voice quality features, such as creak or whisper, are left out of consideration. Nevertheless, it could be argued that these types of information can be relevant to improve the segmental quality of synthetic speech, especially with respect to acceptability (naturalness, pleasantness).

If one wants to go further than improving synthetic phoneme quality from a purely functional point of view, i.e. in terms of identification as the intended phoneme, one may consider taking recourse to highly trained listeners who have an extensive symbol inventory at their disposal to denote subtle and deviant sound characteristics, without any preimposed restrictions. The possibilities of this approach were first explored by Van Gerwen and Vieregge [5], who used the narrow transcriptions made by one experienced ear-phonetician to improve the quality of a text-to-speech conversion system for Spanish. More than 200 words were transcribed twice, the first time to assess segmental imperfections, the second time to check the effects of alterations.

The present study was designed to gain insight into the relative merits of narrow transcriptions and data yielded by an open response identification task as means of diagnostically evaluating the segmental quality of synthetic speech. The comparison took place within the framework of the Dutch SPIN-ASSP program (1985-1990), which was set up to improve text-to-speech conversion for Dutch. First, methodological details will be given. Next, results will be presented and discussed.

2. METHOD

2.1 Open response identification task

In April 1990 a segmental intelligibility test was conducted to evaluate the output of seven synthesis systems for Dutch. For each system, 100 CVC words and 100 VCCV words, phonotactically permissible combinations of Dutch phonemes, were presented in an open response identification task. Most words were meaningless, a few were meaningful. Each phoneme was presented in several phonetic contexts (for further details, see [1]). Eleven advanced students of English from the University of Nijmegen served as subjects. All had some practical knowledge of phonetics, specifically applied to the pronunciation of English, but none had any experience in listening to synthetic speech. They were paid for their participation.

Each CVC and VCCV stimulus word was presented once, with an interstimulus interval of 4 sec. The responses were typed on terminal keyboards. All consonants and vowels had to be identified, using a specially developed, simple but unambiguous notation system. The task was an open response task in the sense that any combination of phonemes could be responded with, provided the number of phoneme responses corresponded with the number of intended phonemes in the stimulus word. At a later stage, the subjects' responses were analyzed (semi-automatically) in terms of percentages correct phoneme identification and phoneme confusions.

The identification task proper was preceded by a short training of 30 minutes in which the notation to be used was explained and practiced. Furthermore, in the actual identification task, each subblock of CVC and VCCV stimuli was preceded by 10 practice stimuli of the corresponding type and synthesis system.

2.2 Transcription task

A large part of the stimulus material presented in the identification task was transcribed by 30 students of Speech and Language Pathology from the University of Nijmegen as part of a comprehensive course in segmental transcription of pathological speech. They worked in pairs, each of the 15 pairs yielding consensus transcriptions for 70 CVC and VCCV words, 10 for each synthesis system.

Since it would have been too time-consuming to examine the transcriptions of all phoneme realizations, it was clear a selection had to be made for the purpose of the present study. It was decided to consider the transcriptions of the realizations of one target phoneme for each of the seven CVC and VCCV phoneme positions for each of the seven synthesis systems, i.e. the realizations of 49 target phonemes. In view of the special relevance of a good diagnosis for poor phoneme realizations, in each case the phoneme which had yielded the lowest mean intelligibility score in the identification task was selected. The intelligibility scores for the target phonemes varied considerably (between 0% and 84% correct), as a function of phoneme category (vowel versus consonant), phoneme position, and synthesis system.

On the average, each target phoneme occurred in 5.9 different words, amounting to a total of 291 phoneme realizations. The students' consensus transcriptions of these phoneme realizations were checked by the second author, an earphonetician experienced both in the transcription of normal and pathological speech. A small part of the material (about 15%) was transcribed by him alone. The transcription system used was the one described in [4], i.e. the Extensions to the International Phonetic Alphabet for the transcription of atypical speech.

3. RESULTS AND DISCUSSION

The neatest way to establish the relative merits of an identification test and transcription as tools for improving synthetic speech would be a pretest-posttest design in which the effects of alterations based on the outcomes of the two methods were independently assessed and compared. It may be clear that this approach is practically unfeasible.

Instead, we decided to use the results from the identification task as a reference for establishing the possible usefulness of transcription as an alternative means in diagnostic evaluation. After all, synthetic speech is primarily developed to allow man-machine communication in various applications. So, a first prerequisite of synthetic output is that it can be understood by "normal" human listeners, that the sounds produced are interpreted in terms of the intended phonemes. Any segmental diagnostic evaluation method should be capable of showing to what extent this basic condition is fulfilled. In other words, if transcription is to be considered as a valid diagnostic tool the data it yields should agree with the identification results obtained in a segmental intelligibility test.

Ideally, in addition to this basic information, narrow transcriptions should yield more. However, as was stated before, the usefulness of this extra information for diagnostic purposes can really only be assessed by formally testing the perceptual effects of the resulting alterations applied to the system in question. In the present study all transcription details throwing light on particular synthesis characteristics were considered as potentially useful on two conditions, (1) that they were systematic, i.e. occurred in at least half of the transcriptions pertaining to the realizations of one particular target phoneme, and (2) that they could not be inferred from the results yielded by the identification task.

With these definitions of what constitutes basic and extra information in mind, the transcriptions were carefully examined. To facilitate generalizations, each series of transcriptions pertaining to the realizations of the same target phoneme were assigned to one of the following three categories:

1. Equivalent to the identification method, i.e. leading to the same qualitative and quantitative interpretation in terms of correct and incorrect phonemes.
2. More informative, leading to the same qualitative and quantitative interpretation and, in addition, providing extra information as defined above.
3. Misleading, leading to a qualitatively or quantitatively different interpretation, suggesting an overestimation or an underestimation of phoneme intelligibility.

The distribution of the (series of) transcriptions for the 49 target phonemes was 30, 7, and 12 in categories 1, 2, and 3, respectively. So, in 30 cases (61%), spread over all 7 synthesis systems, the transcription and identification methods were found to be equivalent in the sense that they yielded the same basic information in terms of correct and incorrect phonemes.

In 7 cases (14%), spread over 5 systems, transcription appeared to be more informative, providing additional information which was considered potentially useful for the improvement of the segmental quality of the synthesis system at hand. The information pertained to voice quality (3 cases), to the undue presence of a final consonant in VCCV words (2 cases). to diphthongization (1 case), and to overly strong phoneme realization (1 case).

In 12 cases (24%), spread over 6 systems, the transcriptions proved misleading in the sense that they did not correspond with the pattern of responses obtained in the identification task. In 7 cases the difference was qualitative, in 5 cases quantitative. Of the latter, 2 would have led to an overestimation and 3 to an underestimation of phoneme intelligibility. We were somewhat amazed by the relatively high number of category 3 cases, since we had expected the transcriptions to generally show the same phoneme distribution as found in the identification task. The point was not clarified by an inspection of the original, unchecked transcriptions, since the differences found hardly affected the categorization (there was only one doubtful case).

In any case, the outcome of the present study suggests that it is somewhat risky to use narrow transcriptions made by highly trained listeners as a substitute for an open response identification task with moderately trained listeners. Apparently, the transcriptions are not always a good predictor of the communicative adequacy of a system in terms of phoneme categorization. Moreover, the transcription approach has other disadvantages as well. One needs highly skilled listeners who have been trained extensively; the method is extremely time-consuming; the designer of the synthesis system has to be able to interpret the transcription symbols; and the data are very difficult to summarize in an insightful manner.

This does not mean to say that we deny any role to transcription in the evaluation of synthetic speech. After all, the present study revealed several cases where transcriptions provided potentially useful diagnostic information not deducible from the results yielded by an open response identification test. The reader may recall that only those transcription details were categorized as potentially useful that occurred systematically in the transcriptions of the realizations of the same target phoneme. This is a rather strict condition, and it cannot be excluded that much more potentially useful information was contained in the transcriptions of individual items.

We are convinced that narrow transcription can contribute significantly to the improvement of synthetic speech if it is used with specific questions in mind, i.e. at a more "local" level. One could think, for example, of a configuration in which a system developer consults one or more transcribers to test the validity of specific hypotheses based on his own perception - after all, it is a well-known fact that system developers generally lose objectivity when listening to the output of their own system - or, perhaps even better, to clarify the outcomes of a formal identification test. In our experience, the efficiency of this procedure is enhanced if the written transcriptions are accompanied by oral explanations.

REFERENCES

[1] BEZOOIJEN, R. VAN (1990), *Evaluation of speech synthesis for Dutch: comparison of synthesis systems, intelligibility tests, and scaling methods*, SPIN-ASSP Report no. 22, Foundation for Speech Technology, Utrecht.

[2] BEZOOIJEN, R. VAN & POLS, L.C.W. (1987), "Evaluation of two synthesis-by-rule systems for Dutch", *Proc. Eur. Conf. Speech Techn.*, Edinburgh, 1, 183-186.

[3] CARLSON, R. & GRANSTROM, B. (1989), "Evaluation and development of KTH text-to-speech system on the segmental level", *Proc. ESCA Workshop Speech Input/Output Assessm. and Speech Databases*, Noordwijkerhout, 1.3.1-1.3.4.

[4] DUCKWORTH, M., ALLEN, G., HARDCASTLE, W., & BALL, M. (1990), "Extensions to the International Phonetic Alphabet for the transcription of atypical speech", *Clinical linguistics & phonetics, 4*, 273-280.

[5] GERWEN, R.P.M.W. VAN & VIEREGGE, W.H. (1989), "Evaluation of an automatic text-to-speech conversion system for Spanish", *Proc. Workshop Speech Input/Output Assessm. and Speech Databases*, Noordwijkerhout, 3.5.1-3.5.4.

[6] HOUSE, A.S., WILLIAMS, C.E., HECKER, M.H., & KRYTER, K.D. (1965), "Articulation-testing methods: consonantal differentiation with a closed response set", *JASA, 37*, 158-166.

[7] POLS, L.C.W., LEFEVRE, J.P., BOXELAAR, G., & SON, N.E. VAN (1987), "Word intelligibility of a rule synthesis system for French", *Proc. Eur. Conf. Speech Techn.*, Edinburgh, 1, 179-182.