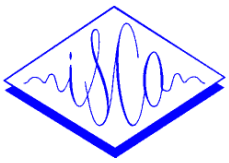

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

TEXT-TO-SPEECH CONVERSION FOR DUTCH: COMPREHENSIBILITY AND ACCEPTABILITY

Willy Jongenburger* and Renée van Bezooijen**

* Institute of Language and Speech, Phonetics Section
** Institute of General Linguistics and Dialectology
University of Nijmegen, P.O.Box 9103, 6500 HD Nijmegen, The Netherlands

ABSTRACT

In the majority of studies to date, evaluation of synthetic speech has been focussed on segmental intelligibility. However, as the development of text-to-speech systems progresses and real-life applications are getting more common, the need for evaluation at higher levels of linguistic organization becomes increasingly relevant. In the present paper two studies are reported evaluating text-to-speech conversion for Dutch at the level of the paragraph. The first study determined comprehensibility for visually impaired and sighted subjects. The second study examined acceptability of voice-and-speech aspects of synthetic output as a function of experience.

1. GENERAL INTRODUCTION

In the majority of studies to date, evaluation of automatic text-to-speech conversion has been confined to assessing segmental quality, either by using one of the well-known rhyme tests or by using an open response phoneme identification test. In these tests the stimuli typically consist of short meaningful or meaningless words with a predefined phoneme structure, mostly C(onsonant) V(owel) C(onsonant). Although reasonable segmental quality is an essential prerequisite for text-to-speech output to be comprehensible and acceptable at higher levels of linguistic organization, it is not a sufficient condition.

While the importance of other aspects of the text-to-speech conversion process (e.g. correct grapheme-to-phoneme conversion, accurate assignment of word stress, appropriate intonation, natural voice quality, etc.) has been recognized for some time, the number of tests related to these features are relatively scarce. Also, comprehensibility of texts consisting of several sentences, as opposed to intelligibility at the word level, has been unduly neglected (for an overview of methodological aspects of the evaluation of text-to-speech systems, see [2]).

In the present contribution two studies are reported that pertain to the evaluation of text-to-speech conversion for Dutch at the paragraph level. The first study aimed at assessing comprehensibility, i.e. the extent to which listeners are able to extract content related information from synthesized messages. This study, in which both visually impaired and sighted subjects participated, was carried out in the final stage of the Dutch SPIN-ASSP program, a five-year project (1985-1990) involving basic and applied research in the area of advanced text-to-speech systems [1]. The second study evaluated the acceptability of output characteristics in terms, for example, of subjectively perceived accuracy of pronunciation, adequacy of word stress, liveliness, and pleasantness of voice. We were specifically interested in the development of acceptability as a function of experience with text-to-speech systems. This study was carried out within the framework of the ELK-project, involving the introduction of an electronic newspaper in the Netherlands with the aim of facilitating the regular and direct access of visually impaired people to up-to-date news. Subjects were recruited from participants in the experimental phase of the ELK-project.

2. COMPREHENSIBILITY

2.1 Introduction

As mentioned above, the comprehensibility study aimed at assessing to what extent sighted and visually impaired listeners are able to extract content related information from synthesized texts. This study was carried out with a view to the growing market for speech technology applications, in this case related to automatic text-to-speech conversion. Applications can be found both in the field of aids for the (visually) handicapped and in information services for the non-handicapped. Examples of the latter are text-to-speech converted weather forecasts and traffic broadcasts. An example of the former are electronic newspapers (see section 3.1).

2.2 Method

The speech material considered in the comprehensibility study consisted of eight slightly modified articles taken from a national Dutch newspaper. There were two parallel series (A and B) of texts, each time two pertaining to the weather (1A and 1B), nature (2A and 2B), disaster (3A and 3B), and small event (4A and 4B). The number of sentences per text was 3, 4, 5, or 6 and the mean number of words was 59, ranging from 30 in the shortest text to 85 in the longest. The eight texts were realized and presented to groups of listeners in three versions:

1. "Natural", i.e. spoken by an experienced male reader who had been involved in reading texts for the blind for 18 years.
2. "TTSAuto" (automatic text-to-speech conversion), i.e. realized by means of standard Dutch diphone-based text-to-speech conversion, using an inventory comprising both unreduced diphones, segmented from stressed syllables, and reduced diphones, taken from unstressed syllables. Except for some minor modifications, this version represents the present state-of-the-art of diphone-based text-to-speech conversion as developed at the Institute of Perception Research (IPO) in Eindhoven, The Netherlands.
3. "TTSCorr" (corrected text-to-speech conversion), i.e. realized in the same manner as TTSAuto, but corrected manually. Correction concerned errors in word stress and grapheme-to-phoneme conversion. Moreover, the number of vowels realized with an unreduced diphone was augmented, certain glottal stops were removed, and extra degemination and assimilation was applied. Imperfections in the diphones or diphone concatenation, the location of sentence accents, or durational characteristics were not amended. This version is a realistic construction of the (expected) future state-of-the-art.

The 8 (texts) x 3 (versions) = 24 text realizations were presented to groups of visually impaired (N=16) and sighted subjects (N=24). All subjects participated in the experiment on a voluntary basis. The visually impaired were selected at random from among the subscribers to a center distributing spoken magazines and books (the CGL in Grave, The Netherlands). The two groups differed in several respects: sex, age, and educational level. The group of visually impaired consisted of 7 males and 9 females, whereas the group of sighted consisted of 3 males and 21 females. The mean age of the visually impaired was 40, ranging from 14 to 79; the mean age of the sighted was 22, ranging from 22 to 27. The level of education of the visually impaired ranged from primary education to tertiary education, whereas all sighted subjects were university students. The characteristics of the visually impaired subjects are representative of the future subscribers to the electronic newspaper. The sighted subjects constitute a representative sample of the population of students from the Faculty of Arts.

In both experiments comprehensibility was assessed by means of the same texts and the same questions. However, due to factors not relevant to the present study, there were some differences in the way in which subjects were distributed over conditions. Also, in the case of the sighted subjects, stimuli were presented group-wise over earphones. The questions were presented in written

10.21437/ICSLP.1992-134

form, after presentation of each text, and the subjects had to answer the questions, also in written form, within a prefixed period of time (10 sec per question). In the case of the visually impaired subjects, the test was run individually at their homes. No earphones were used. The questions were presented orally, after presentation of each text, and the subjects answered the questions orally, with no time limitations. The answers were taperecorded to be transcribed at a later stage.

2.3 Results and discussion

In Table 1 the percentages correctly answered questions for each text are presented, separately for the three text versions and the two subject groups. The texts are ordered according to length, from short to long. It must be noted that answers were only counted completely correct if all elements in the text which we thought relevant to a question were contained in the answer. If not, quarter or half points were subtracted. So, the criteria applied to determine correctness were rather severe and strict.

It can be seen from Table 1 that the percentages correct answers are somewhat higher for the sighted subjects than for the visually impaired. This holds for all three text versions. In order to test the significance of the differences found, the scores were subjected to an analysis of variance with the fixed factors of group (2 levels) and version (3 levels) and the random factor of text (8 levels). The level of significance was set at 5%. The factor "group" was not significant, nor the interaction "group x version". From this it has to be concluded that statistically speaking the two subject groups performed equally well.

As mentioned above, the visually impaired subjects were older and less educated than the sighted subjects. In view of the nature of the task, with an important memory and cognition component, these factors could be hypothesized to favor the sighted subjects, just like the use of earphones during the listening task. Possible factors in favor of the visually impaired subjects are the fact that there were no limitations on their response time. Moreover, the visually impaired subjects may have been more motivated, synthetic speech opening up new possibilities of access to extensive up-to-date information services. Also, visually impaired people may benefit from their greater experience with auditory information processing; much information processed by sighted people is accessed via the visual channel.

The fact that the performance of the two subject groups was found to be similar can be explained in different ways. First, the hypotheses concerning the differential influence of various factors mentioned above may be invalid. Second, various factors may have canceled each other out. For example, the influence of the younger age and higher educational level of the sighted subjects may have been compensated for by the greater motivation and/or experience with auditorily presented information on the part of the visually impaired. In any case, the results of the present study suggest that comprehensibility of natural and synthesized texts does not have to be tested separately for sighted and non-sighted people, results for the one group being generalizable to the other.

As the data in Table 1 further suggest, there was a significant effect of the factor "version". To examine the precise nature of this effect, an a posteriori multiple comparison test (Tukey's HSD) was carried out. Only one pair of versions was found to differ significantly ($p<.05$) from each other, namely TTSAuto and Natural. For both subject groups the automatic text-to-speech realizations are the most difficult to comprehend, whereas the human produced text realisations are the easiest. The manually corrected text-to-speech realisations occupy an intermediate position. This is in line with what one would expect. However, it is worth noting that even in the TTSAuto condition about two thirds of the open content questions were answered correctly, in both groups.

We finally would like to point out that comprehensibility seems to vary as a function of text, texts 1A and 1B, the weather forecasts, yielding the highest numbers of correct answers. These two texts A first explanation of their higher comprehensibility could be sought in their shorter length. The shorter the text, the smaller the memory demands placed on the subjects, the easier correct answering of the questions.

Further examination of the data in Table 1 suggests that, however plausible, text length is probably not the only factor playing a role, since otherwise one would expect a steadily decreasing number of correct answers going from texts 1A and 1B to texts 4A and 4B. However, texts 4A and 4B generally appear to be more comprehensible than texts 2A, 2B, 3A, and 3B. Since the rank order of comprehensibility seems to be independent of text version, particular text related text-to-speech conversion problems can be ruled out as an explanation for the differences found among the texts. Additional factors affecting comprehensibility may relate to the predictability of the structure and the content of the texts (weather forecasts being highly standardized and therefore highly predictable in all respects), shared interest in the topic on the part of the subjects, and the information density of the texts, in terms, for example, of the ratio of content words and function words, the ratio of "old" information and "new" information, syntactic complexity, or sentence length. These factors are worth to be further looked into.

Table 1. Mean percentages correct answers for 8 texts, separately for the 3 text versions and the 2 subject groups. Nat = natural.

	Sighted			Visually impaired		
Text	TTSAuto	TTSCorr	Nat	TTSAuto	TTSCorr	Nat
1A	90	96	98	83	97	90
1B	85	88	100	100	93	97
2A	77	80	84	48	70	80
2B	68	73	70	52	70	75
3A	65	69	71	60	40	70
3B	46	62	67	33	80	70
4A	71	79	85	58	76	88
4B	72	85	94	82	88	80
	--	--	--	--	--	--
Mean	72	79	84	64	76	81

3. ACCEPTABILITY

3.1 Introduction

As mentioned in section 1, the acceptability study aimed at evaluating the output characteristics of text-to-speech conversion for Dutch at the paragraph level. The evaluation took place within the framework of the ELK-project, which involves the introduction of an electronic newspaper in the Netherlands for the visually impaired. The technique, which has been positively received in Sweden [3,4], is based on the transmission of a digitized newspaper, which is subsequently stored in the PC of the user. By means of a text-to-speech system (and/or braille) combined with a search system, the visually impaired can "read" the newspaper. Transmission takes place at night, at the same time that the newspaper is printed.

Until December 1992 the ELK-project is in an experimental phase, in which participate 5 deaf and visually impaired adults who read the newspaper in braille and 45 visually impaired adults who read the paper by means of a text-to-speech system. During the initial stage, two different text-to-speech systems are used concurrently, the one developed at the Institute for Perception Research in The Netherlands ("the IPO system", see [5]), the other ("the Apollo system") produced by Dolphin Systems in the United Kingdom and distributed in the Netherlands by CIG ("Centrum Informatica voor Gehandicapten"). The two systems differ in several respects, in price (the IPO system being more expensive), in hardware (e.g. type of synthesizer), and in software (e.g. the grapheme-to-phoneme and word stress assignment components being more sophisticated in the IPO system). Moreover, the IPO system is diphone based, whereas the Apollo is allophone based. On the whole, the IPO system may be said to be of a higher quality than the Apollo. On the other hand, the Apollo is more flexible in the sense that the user has more options, e.g. with respect to pitch height, voice quality, and tempo.

The two systems were distributed randomly over the 45 visually impaired taking part in the experimental phase of the ELK-project. From the two groups of users, a representative sample of subjects was selected to participate in the evaluation described in this paper. The aim of the evaluation was fourfold:

1. to determine the acceptability of the two text-to-speech systems
2. to determine the intelligibility of the two text-to-speech systems
3. to gain insight into the attitudes towards the electronic newspaper
4. to determine the adequacy of the search system

The present paper reports on a study related to the first aim. The general aim of the acceptability study was to assess subjectively perceived output characteristics of the two text-to-speech systems as a function of experience. First, we were interested in the development of the evaluative reactions of the two user groups towards their own system. Does experience with a particular text-to-speech system have a positive effect on the evaluation of that system, or, on the contrary, do some aspects get more irritating over time? In addition, we were curious to know whether there would be a carry-over effect. Does experience with one particular text-to-speech system affect the evaluation of another text-to-speech system? Of course, we also wanted to assess whether the reactions towards the two systems differed and whether the two user groups were comparable in this respect. So, in short, we aimed at determining to what extent acceptability varied as a function of system, user group, and experience.

3.2 Method

The stimulus material used in the acceptability study consisted of six short weather forecasts with similar content and structure. All texts contained five sentences of between six and eleven words. Three texts were synthesized with the IPO system and three with the Apollo system.

The texts were presented to 24 visually impaired subjects (13 men and 11 women), a subgroup of those taking part in the experimental phase of the ELK-project. Their mean age was 42 years, ranging between 18 and 82. Educational level varied from technical and vocational training for 12-16 years old to academic degree. Nine of the subjects worked with the IPO system, 15 with the Apollo. Most of the subjects had had already some experience with a personal computer prior to the experiment.

In order to study acceptability as a function of experience, the subjects were visited at their homes at three points in time:

t1, at the beginning of the project, before the text-to-speech systems were set up (at this point in time, the division into two user groups, those working with the IPO system and those working with the Apollo, had not yet become effective)

t2, after about a month (between 2.5 and 4.5 weeks) of experience with the reading device

t3, after about two months (between 8 and 10 weeks) of experience with the reading device

Acceptability was assessed by means of ratings on 10-point scales related to ten aspects of voice-and-speech, namely intelligibility, general quality, naturalness, precision of articulation, accuracy of pronunciation, pleasantness of voice, adequacy of word stress, appropriateness of tempo, liveliness, and fluency. The higher the score, the more positive the evaluation. For each aspect, two different taperecorded texts were made audible successively, one synthesized with the IPO system, the other with the Apollo system. So, either text was presented 10 times. At each point in time, different texts were presented to any one subject.

3.3 Results and discussion

As mentioned above, the general aim of the acceptability study was to assess subjectively perceived output characteristics of the IPO- and Apollo systems by two user groups as a function of experience (no experience, 1 month, 2 months). To that end, the ratings for each of the ten voice-and-speech aspects considered were subjected to an analysis of variance with the fixed factors of group (2 levels), system (2 levels), and experience (3 levels). The results of the tests were found to be rather redundant in the sense that many scales showed similar patterns of significant effects. In order to gain insight into the relationships among the ten output characteristics evaluated and reduce the dimensionality of the evaluation space, a factor analysis was carried out on the ratings. Two factors emerged with an eigenvalue >1. Together they accounted for 61% of the variance in the original data.

The structure of the varimax rotated factors was rather straightforward, as can be seen in Table 2. The highest loadings (>.70) on the first factor are from intelligibility, general quality, and precision of articulation, suggesting an interpretation in terms of segmental quality. The scale "intelligibility" was taken to best represent this factor. The variables with the highest loadings on the second factor are naturalness, pleasantness of voice, and adequacy of word stress. This factor seems to be represented most adequately by the scale "naturalness". So, intelligibility and naturalness were taken to be the two central dimensions underlying the evaluative reactions.

Table 2. Rotated factor matrix

	Factor 1	Factor 2
Intelligibility	.84	.18
General quality	.78	.30
Naturalness	.28	.79
Precision of articulation	.79	.26
Accuracy of pronunciation	.40	.40
Pleasantness of voice	.47	.70
Adequacy of word stress	.05	.86
Appropriateness of tempo	.40	.50
Liveliness	.40	.63
Fluency	.68	.25

In Figure 1, the results for perceived intelligibility are depicted graphically. On the left the mean ratings given by the subjects using (or designated to use) the IPO system are given, on the right the mean ratings for the subjects using (or designated to use) the Apollo can be found. The analysis of variance revealed two significant main effects, namely of the factors "system" and "experience", and three significant interactions, namely of group x system, of system x experience, and of group x system x experience. An a posteriori multiple range test (Student-Newman-Keuls procedure) was carried out to gain insight into the precise nature of the effect of the factor "experience" and its interactions with the other two factors. No significant contrasts were found between t1, t2, and t3 for those working with the IPO system, for neither of the two systems. As for those working with the Apollo, one significant contrast was found, namely between the perceived intelligibility of the Apollo at t1, i.e., when the subjects were confronted with the output of this system for the first time, and at t2 and t3, after one month and two months of experience with the Apollo system.

The main findings with respect to intelligibility can be summarized as follows. Overall, the IPO system was perceived as being more intelligible than the Apollo (mean ratings of 7.5 and 6.5, respectively, averaged over the two user groups) and perceived intelligibility increased as more experience was gained (6.6, 7.2, 7.2 at t1, t2, and t3, respectively, averaged over the two systems). However, as shown by the results of the statistical tests and the data presented in Figure 1, the picture is more complex than that.

As for those working with the IPO system, there is no development in time; their own system is assigned uniformly high intelligibility ratings at all three measuring points (7.8, 8.1, 8.2 at t1, t2, and t3, respectively), whereas the Apollo system is assigned uniformly low intelligibility ratings (6.0, 6.1, 6.2). Apparently, exposure to the high-quality output of the IPO system does not raise perceived intelligibility. On the other hand, it cannot be excluded that the absence of an observed increase is due to a ceiling effect, in the sense that the subjects reserved points 9 and 10 on the 10-point rating scale for intelligibility of human produced speech.

In addition, it can be observed that for the users of the IPO system the difference in perceived intelligibility between the two systems remains constant. This suggests that there is no carry-over effect, i.e., experience with listening to the IPO-output does not have a positive effect on perceived intelligibility of the Apollo output (where there does seem to be some room for higher ratings).

As for the Apollo group, the situation at t1, prior to experience with Apollo output, seems to be comparable to the differential evaluation at t1 of the subjects designated to work with the IPO system; the IPO system is assigned a relatively high intelligibility score (7.2), whereas the Apollo is given a relatively low intelligibility score (5.5). In fact, at this point in time one would indeed expect the two groups of users to have similar reactions towards the two systems, since both of them are presented with synthetic speech for the first time.

However, at t2, after one month of exposure to the Apollo, the situation has changed drastically. The output of the Apollo is perceived to be much more intelligible now, equally intelligible in fact as the IPO system. This finding strongly suggests that listeners are indeed able to learn to interpret the segmental characteristics of a particular system, to discover the systematics underlying their production. Apparently, this learning process is finished within a month; after that, no further increase takes place. So, in this case, where the subjects received a system with a relatively low segmen-

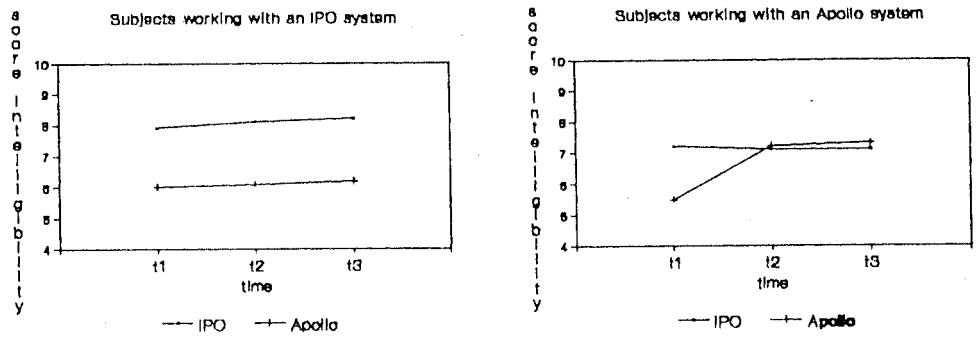

Figure 1. Intelligibility scores of two user groups for two text-to-speech systems at three points in time.

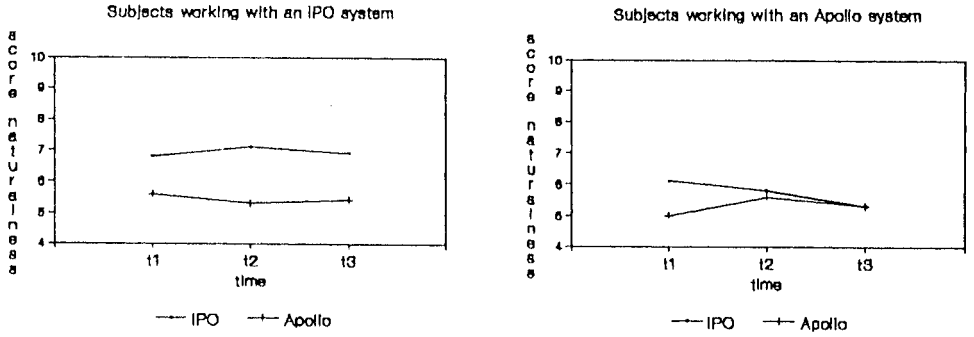

Figure 2. Naturalness scores of two user groups for two text-to-speech systems at three points in time.

tal quality, experience did indeed have a positive effect on the evaluation (of course, it could also be that now there was some room for improvement to manifest itself in the ratings). But again, just like for the group of IPO-users with respect to the Apollo, there was no carry-over effect, the ratings assigned by the Apollo-users to the IPO system remaining constant over time.

In Figure 2, the results for naturalness are shown. At first sight, the pattern looks quite similar to that shown in Figure 1. And indeed, just like for intelligibility, the factor "system" proved to have a significant effect on the ratings as well as the interaction of group x system. However, there are some differences too: with respect to naturalness, no significant effect was found of experience, nor of the interactions of system x experience and of group x system x experience. As compared to intelligibility, one significant effect was added, namely of the factor "group".

How can the results for naturalness be summarized? Again, it must be concluded that overall the IPO system is evaluated more positively than the Apollo (6.2 and 5.4, respectively, averaged over the two user groups). However, this is largely due to the ratings of the IPO-group and the ratings of the Apollo group at t1. At t2 and t3, those working with the Apollo do not seem to perceive any difference in naturalness between the two systems; the IPO system is given the same, relatively low, naturalness ratings as the Apollo (5.8 and 5.6, respectively, at t2 and 5.3 and 5.3, respectively, at t3). It is somewhat difficult to interpret these low naturalness ratings for the IPO system. It seems to suggest a carry-over effect in the sense that the negative attitude towards the Apollo, i.e., the system the subjects worked with, brought about a negative effect towards synthetic speech in general. Anyway, it is the low ratings of the Apollo users for the IPO system at t2 and t3 that explain the significant effect of the factor "group".

Finally, it must be noted that, in contrast to what was found for intelligibility, exposure to the Apollo did not lead to a more positive evaluation of naturalness at t2 and t3 as compared to t1. This points to a fundamental difference between the two aspects of synthetic output. Whereas (perceived) intelligibility can increase through learning, perceived naturalness would have to increase because of habituation, in this experiment specifically to inadequate word stresses and the unpleasantness of the voice (see Table 2). Apparently these voice-and-speech characteristics are not something one gets used to; if they are evaluated negatively in a first confrontation, this remains so, even after repeated exposure.

ACKNOWLEDGMENT

The comprehensibility study was supported by the Foundation for Speech Technology, which is funded by the Dutch National Program for the Advancement of Information Technology (SPIN). The acceptability study was funded by the Dutch Ministry of WVC, "het Nederlandse Revalidatiefonds" and "het Preventiefonds".

REFERENCES

[1] S.G. Nooteboom and R. van Bezooijen. "Five Years of Coordinated Research on Text-to-speech Conversion for Dutch: an Overview." *Twelfth International Congress of Phonetic Sciences*, Aix-en-Provence, Vol.3, pp. 470-473, 1991.

[2] R. van Bezooijen and L.C.W. Pols. "Evaluating Text-to-speech Systems: some Methodological Aspects."*Speech Communication*, Vol. 9, pp. 263-270, 1990.

[3] E. Hjelmquist. "Daily Newspapers for Visually Handicapped." *Scandinavian Public Library Quarterly*, Vol.1, pp. 20-24, 1988.

[4] E. Hjelmquist, B. Jansson, and G. Torell, G. "Blind Persons' Reading of Daily Newspapers with Computer Mediated Technique." *Internal Report Department of Psychology*, Göteborg University, 1989.

[5] R.J.H. Deliège. "A Stand-alone Text-to-speech System." In V.J. van Heuven and L.C.W. Pols (eds.): *Analysis and Synthesis Speech, strategic research towards high quality text-to-speech generation*. Berlin: Mouton de Gruyter, 1992 (in press).