Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 23 (1999), 59-65

THE INTELLIGIBILITY OF NAMES

Judith Rog, Angelien Sanderman[1], and Louis Pols

Abstract

Naturally spoken names are frequently hard to understand if one does not know beforehand what to expect. This might be even worse for names generated by a text-to-speech synthesizer. Still, it would be very advantageous for various telecommunication applications if such a service was available. In this M.A. thesis project the intelligibility of various name types (surnames, company names, street names, and city names), spoken naturally and via two different synthesizers, whether or not using a pronunciation lexicon, was systematically evaluated. Error types were studied, various scoring methods were developed, and phoneme confusions were evaluated. Various recommendations are done for future applications.

1. Introduction

This short contribution is a summary of the M.A. thesis of the first author. Her work has been executed at KPN Research (the research and development organisation of KPN, a Dutch telecommunications operator), under the supervision of the second and third author. For more details see the full report (Rog, 1999).

The overall quality of Dutch TTS (text-to-speech) systems is not too bad nowadays. The intelligibility is actually quite good when the listener has some idea about what is going to be said or when s/he is familiar with the speech. However, when this is not the case, intelligibility decreases.

If we look at name pronunciation as a special type of speech, we find that the intelligibility of names, synthesized by a TTS system, is generally worse than that of normal speech. There are several reasons for this. Some of them are due to mistakes in the so called grapheme-to-phoneme conversion part of the TTS system:

1. Word stress in Dutch names is often placed on another syllable (namely the first) than predicted from the pronunciation rules of normal Dutch words. Take for example the word/name: goedkoop/Goedkoop. The word is pronounced as /xut-'kop/, the name as /'xut-kop/;
2. There are often traces of archaic spelling in names that are no longer used in normal words anymore. Take for example the name Baers. Normal pronunciation rules would predict /ba-'Ers/, whereas in Dutch it should be /'bars/;
3. Finally there are a lot of foreign names which do not hold to the Dutch pronunciation rules.

[1] KPN Research, Leidschendam

Furthermore, names are also hard to understand, because:

1. There is usually little or no help from the context;
2. There are many names that are unfamiliar and one might hear them for the first time;
3. Every phonologically correct sequence can be a name;
4. There are a lot of names that differ only in one sound from another name.

In several services that are supplied by telecommunication companies, or will be supplied in the future, names play an important role. With the use of speech synthesis the costs of such services could be reduced. To study this further, two evaluations of the overall quality and intelligibility of Dutch TTS systems were performed (Rietveld et al., 1997; Sluijter et al., 1998). From these two evaluations it could be concluded that more research had to be done on the specific topic of names. Thus another (unpublished) experiment was performed in which the quality and intelligibility of names synthesized by two Dutch TTS systems was evaluated. Both systems were tested under two conditions. Once *without* the use of an extra lexicon and once *with* the use of an extra lexicon of names. ONOMASTICA (1995), a pronunciation dictionary of a large number of European names, was used for this purpose. The two systems that were tested at that time will be referred to as System 1 (a Dutch TTS system that makes use of diphones and LPC) and System 2 (a Dutch TTS system that makes use of diphones and MBROLA). The various names that were used for testing consisted of four *types* of names: surnames, company names, street names, and city names. The global results are presented in Table 1. These average percentage scores are derived from manual 0-1 judgements per word (1 = correct apart from spelling errors, 0.5 = partially correct, and 0 = incorrect).

Table 1. Average word intelligibility for 2 TTS systems with or without using a lexicon.

	without lexicon	with lexicon
System 1	39%	52%
System 2	62%	64%

Analyzing the data furthermore showed that there was no significant difference in intelligibility score between the names if they were split in six groups sorted by frequency of occurrence. The four types of names did show different intelligibility scores and so did the two groups related to using (or not) the extra lexicon. The type of name had more influence on the names produced by System 1 than it had on the names of System 2. Whether the lexicon was used or not had also more influence on the intelligibility scores of System 1 than on the scores of System 2, as can be seen in Table 1. This effect might be related to the fact that some mistakes were made when generating the sound files of the System 2 with the use of the lexicon. As a result of these mistakes the majority of names in this condition were actually realized without word stress or had the word stress on the wrong syllable. This meant that the names in this condition were often not realized in the best possible way, as specified in the pronunciation lexicon. This could explain why for System 2 the difference between the conditions with and without using the lexicon was rather small.

This experiment left some questions unanswered. What does an intelligibility score of 64% mean? Can we be satisfied and lean back with a score of 64%, or should we aim for a higher score, maybe even for 100%?

To answer these and other questions we ran our own experiment in which a comparison was made with an ideal standard, being natural speech. Also further research into what specific elements actually influence the intelligibility of names was done. To this end, first of all, a diagnostic evaluation was performed on the data from the previous experiment. Furthermore, two new scoring methods to score the intelligibility of names were tested in an experiment. In the same experiment the intelligibility of names in natural speech was compared to the intelligibility of synthesized names.

2. Diagnostic evaluation

A diagnostic evaluation was performed on the data from the previous experiment. This was done to see if a certain pattern could be found in the mistakes that people make by trying to understand names. If certain mistakes are made more often than others, this information can be used to improve the TTS systems. The following results were found:

1. Most mistakes were substitution, i.e. one sound was mistaken for another sound;
2. Somewhat to our surprise, most sounds that were misunderstood were sounds in stressed syllables. A possible explanation lies in the fact that in names the first syllable is often the stressed syllable. The first syllable is less predictable than the following ones, which makes the sounds in that (often stressed) first syllable less predictable than those in the following (often unstressed) syllables;
3. Sounds were labelled as being either initial, medial, or final. Initial sounds appeared to be harder to understand than medial or final ones. This finding supports the second point;
4. Consonants were harder to understand than vowels. Vowels are usually louder and are spoken in a more pronounced way than consonants. Furthermore, the group of consonants is bigger than that of vowels;
5. A consonant could either be single or in a position surrounded by one or more other consonants (cluster). Single consonants appeared to be harder to understand than the ones in a cluster. There are several possible reasons for that. If there are more clusters in the non-initial syllable of names, then, according to point two and three above, the consonants in clusters are more predictable than the single ones. It could also be the case that consonant clusters in names (for example the /str/ in 'straat') are in general more predictable than single consonants. Also if the first sound of a cluster is identified and the listener knows that the next sound was also a consonant, the predictability of that consonant increases. For example if /d/ was the first sound and it is followed by another consonant, than this can only be either /w/ or /r/.

Furthermore it was found that street names and company names are harder to understand than place names and surnames. Foreign names and non-Dutch names (for example names with traces of foreign spelling) are harder to understand than Dutch names. Shorter names are easier to understand than longer ones. Names with a high frequency of occurrence are easier to understand than those with a lower frequency.

3. Scoring methods

An important issue when measuring intelligibility is how to define the intelligibility score. When is a name identified correctly and when is it not? If a persons hears the name 'Clinton' and writes down 'Klintyn', what score do we give? Certain sounds can be written down in different ways. So a different spelling does not necessarily mean that the name was not understood. Keeping this in mind, a trained person can score all the responses manually, but this is very time consuming. An automatic or semi-automatic scoring method would therefore be very useful.

An experiment was performed in which three scoring methods were compared with each other and with the scoring used in the previous experiment.

The first method was comparable to the scoring method used in the previous experiment. All responses were scored manually on a three-point scale. The scores depended on how well the first author thought the names were understood. As a guidance some rules were used as to the amount of mistakes that could be made in a name to judge it either as 1, 2 or 3.

The second method was based on a scoring method that had been used at the ESCA Workshop on Speech Synthesis (van Santen et al., 1998). Once the subjects had heard a name, they were asked to type this name on an electronic form on a PC. After this, they were confronted with the name they had heard in the correct orthographic form next to their own answer. They were then asked to make a judgement about their own response. They were asked: "how well do you think you understood the name?" Answers could be given on a three-point scale: 1. Reasonably well, 2. Mediocre, 3. Poor. It was made clear to them that certain sounds could be spelled in different ways. So another spelling would not necessarily imply an incorrect understanding. If the subjects would actually be able to score their own responses in this way, it would reflect a very useful intelligibility score because it is after all the judgement of the users or customers that matters.

The third scoring method makes use of the computer program EVAL that was written by the first author. The program makes use of two already existing programs. One grapheme-to-phoneme (letter-to-sound) program and one DTW (Dynamic Time warping) program. In EVAL the subjects' responses, that are in orthographic form, are converted to a phonetic form using the GRAFON program (Daelemans, 1988). Some adaptations are made to the phonetic forms of the responses and of the stimuli (their phonetic forms are in an ONOMASTICA notation) to make them suitable as input for the next module. In the next module the phonetic form of responses and stimuli are fed into ALIGN[2]. ALIGN is a DTW program. It lines up the two phonetic strings and determines the minimal number of substitutions, deletion or insertions of sounds. On the basis of the number of phonetic features that differ a minimal distance factor is calculated. The score reflects a measure of similarity between the two strings. If this is done for a large amount of stimulus and response combinations, an average distance factor can be calculated per condition. This scoring method EVAL, is performed automatically. The input of the program consists of two lists, one list of responses and another list of the corresponding stimuli. The output consists of a distance factor. This is a very quick way to determine a score.

[2] ALIGN was originally written by Dirk de Vries. A description of the program can be found in Cucchiarini (1995).

4. The experiment

4.1 Set-up

Fifteen subjects were presented with a total of 900 names. All names were preceded by a short carrier phrase. The stimuli were IRS-filtered (300-3400 Hz) to get telephone quality speech. The sample frequency was 8 kHz and an 8 bits resolution was used. The audio files were presented over a headset at about 70 dB SPL. The stimuli were presented to one ear only, as is customary in telephone applications. Once a stimulus was presented the subject could fill in his answer on an electronic form. After that he could click on an OK-button. The response as well as the stimulus would then appear in orthographic form and the subject was then asked to judge, on a three point scale, how well he felt he had understood the name. Every subject was presented with 60 different names.

The same two TTS systems as in the previous experiment were tested. The names were tested under five conditions:

1. Synthesized by System 2 not using the extra lexicon;
2. Synthesized by System 2 using the extra lexicon (ONOMASTICA);
3. Synthesized by System 1 using the extra lexicon;
4. Produced by a male speaker, an operator of the 8008/118 service, the directory enquiry service of KPN;
5. Produced by a female speaker, an operator of the 8008/118 service, the directory enquiry service of KPN.

4.2 Results

For an overall representation of the results, see Figure 1 and 2. The three scoring methods came up with the same ranking order (based on average scores) of the five conditions tested. If we calculate Pearson's correlation we find the lowest correlation between the scores of EVAL and the scores from the previous experiment, to be 0,928* (the * indicates significance at $p < 0.05$). The overall correlation between the raw scores of the three scoring methods used in the new experiment, Kendall's W of concordance, is 0,796*. We can conclude that the three scoring methods were similar and that the subjects can judge their own responses.

In the previous experiment it was found that there was only a small effect of the use of a lexicon on the intelligibility of names as synthesized by System 2. The effect was a lot bigger for System 1. As this was something unexpected, an explanation was sought. As already explained in the introduction, this was linked to the fact that with the generation of the System 2 sound files using the lexicon some mistakes were made. The result was that there were many names in this condition without word stress or with stress on the wrong syllable. It was expected that, if this was corrected, the influence of the lexicon on the intelligibility would become bigger.

However, in the more recent experiment there was still no significant difference between the intelligibility scores of System 2 with or without the use of an extra lexicon.

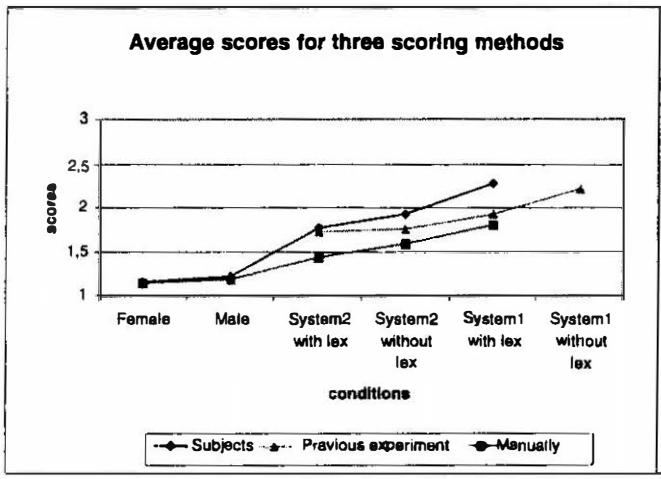

Figure 1: Average scores (according to 3 different scoring methods) for the 5 conditions tested in our experiment and the 4 in the earlier experiment (3 overlapping). The lines between the points do not really exist, but are only added to improve readability.

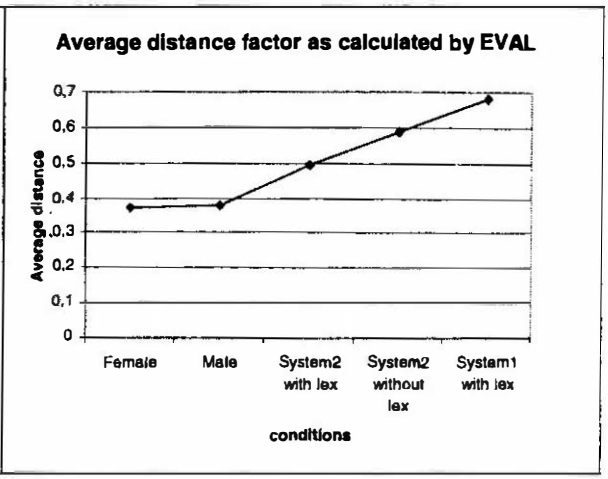

Figure 2: Average distance factor as calculated by EVAL for the 5 conditions that were tested in our experiment. The lines between the points do not really exist, but are only added to improve readability.

There can be several explanations. It could be that the grapheme-to-phoneme rules that System 2 uses are more suitable for names than the ones that System 1 uses. For example the rules of System 2 might put word stress on the first syllable more often. However, that would imply that the rules of System 2 would be less suitable for normal words. This is probably not the case. It could also be possible that there are more names in the internal lexicon of System 2 than of System 1. After a closer look at the internal lexicon of both systems, it is concluded that this not the case either. If we take a look at how the pronunciation of a name changes in a TTS system because of the use of the lexicon, we find that most differences lie only in which syllable is stressed. In the new experiment it was found that the correct realisation of word stress does not have as much influence on the intelligibility of names as was expected.

It was also expected that the names produced in natural speech would be easier to understand than the synthetic ones. However, we also know that even in natural speech names are often hard to understand. If someone introduces himself, his name is often hard to understand the first time, and one often has to hear it a second time in order to get it right. So the question is: are names in natural speech really easier to understand?

In our experiment we had three scoring methods and five conditions. We found that the two natural speech conditions (male and female speaker) are never significantly different from each other. We found that for all three scoring methods natural speech was significantly easier to understand than the following two synthesis conditions: System 2 without the extra lexicon and System 1 with the extra lexicon. The difference between natural speech and System 2 with the extra lexicon was significant for two out of three scoring methods. The third scoring method, EVAL, however showed no significant difference in intelligibility scores between natural speech and System 2 with the use of an extra lexicon. As can also be seen from figure 1 and 2, we can conclude that naturally spoken names are still more intelligible than synthesized names, but the difference is not as big anymore as we had expected.

We also found a significant influence of type of name on the intelligibility of names. Furthermore, if the speech gets less intelligible the judgements of the subjects become more diverse.

5. Conclusions

The final conclusion could be that, whenever a spelling option is available as a fall-back mechanism, the use of TTS systems to synthesize names in telecom services is possible. Even if natural speech would be used, intelligibility would still not be 100%, and a spelling option would still be required. We found that the more intelligible TTS System 2 was closer to natural speech in terms of intelligibility than to the other TTS system tested.

The use of an extra lexicon does not necessarily lead to a higher intelligibility of names. It was found that there are relatively few names that are pronounced 'better' (in terms of perceptually different and more intelligible) because of the use of a lexicon. This does not mean that a lexicon should not be used. It may not necessarily lead to better intelligibility, but our expectation is that it does significantly contribute to the quality of the speech (for example concerning naturalness, voice pleasantness, etc.).

A semi-automatic way of judging the responses of the subjects, EVAL, can be used to determine an intelligibility score.

Subjects can judge their own responses on correctness even if the orthography deviates. This is an important point because these subject are supposed to be representative for the people that would be using the services of KPN in real life.

For the developers of the TTS systems it is strongly recommended to make use of multiphones in the future. It might also be advantageous to leave high-frequent parts of names intact. In this respect one could think of examples like 'straat', 'laan', 'weg', or 'ver' (as in 'Verschuren, or 'Verhoeven').

References

Cucchiarini, C. (1995), "What do transcription agreement indices say about transcription accuracy?", in: K. Elinius & P. Branderud (eds.), *Proceedings of the 8th International Congress of Phonetic Sciences (ICPhS'95)*, Stockholm Sweden, Vol. III, pp. 484-487.

Daelemans, W. (1998), "GRAFON: A grapheme-to-phoneme conversion system for Dutch", *Proceedings of the 12th COLING*, Budapest, 133-138.

ONOMASTICA (1995), *Multi-language pronunciation dictionary of proper names and places*, Final report.

Rietveld, T., Kerkhoff, J., Emons, M.J.W.M., Meijer, E.J., Sanderman A.A. & Sluijter, A.M.C.V. (1997), "Evaluation of speech synthesis systems for Dutch in telecommunication applications in GSM and PSTN networks", *Proceedings of Eurospeech '97*, Rhodes, Greece, 577-580.

Rog, J. (1999), "De verstaanbaarheid van namen" ("The intellegibility of names"), M.A. thesis, Institute of Phonetic Sciences, University of Amsterdam.

Santen, J.P.H. van, Pols L.C.W., Abe M., Kahn D., Keller E. & Vonwiller J. (1998), "Report on the Third ESCA TTS Workshop Evaluation Procedure", *Proceedings of the third ESCA/COCOSDA International Workshop on Speech Synthesis* (Jenolan Caves, Australia), 329-332.

Sluijter, A., Bosgoed, E., Kerkhoff, J., Meier, E., Rietveld, T., Sanderman, A., Swerts, M. & Terken, J. (1998), "Evaluation of speech synthesis systems for Dutch in telecommunication applications", *Proceedings of the third ESCA/COCOSDA International Workshop on Speech Synthesis* (Jenolan Caves, Australia), 213-218.