

On the performance of speech output in a practical setting

E.A.M. Klabbers and R.P.G. Collier

e-mail: klabbers@ipo.tue.nl

Abstract

In spoken dialogue systems, in which humans interact with computers over the telephone, it is essential that the voice output of the system be of high quality. Both the intelligibility and the naturalness of the output should be sufficiently high. There are several techniques for providing a system with speech output, each with its own advantages and disadvantages. This paper discusses a formal evaluation experiment of three speech output techniques. Natural speech was included as a reference condition. The speech was rated on intelligibility and fluency of the output. Additionally, the overall quality of the speech and its suitability for use in a commercial application were assessed.

The results reveal significant differences between the techniques. Diphone synthesis still has an inferior quality compared to the other techniques, both in terms of intelligibility and fluency. Conventional phrase concatenation is quite intelligible, but scores less on fluency. IPO's phrase concatenation is by far the best technique.

Introduction

In spoken dialogue systems, in which humans interact with computers over the telephone, it is essential that the voice output of the system be of high quality. Current spoken dialogue systems produce speech output using different techniques, ranging from fixed pre-recorded messages to fully synthetic speech. Some systems present a small fixed number of sentences, whereas others produce a large but finite number of words to be inserted into a small fixed set of carrier sentences. Some applications even reproduce unrestricted text, for instance newspaper articles. The decision to use a particular speech output technique in an application very much depends on the variability of the information to be presented. Typically, the larger the variability, the smaller the units used for synthesis (ranging from phrases to words to diphones).

The research described in this paper is carried out within the OVIS¹ application domain, which provides train timetable information over the telephone. Its current output typically consists of a small number of carrier sentences with a number of slots in which a large number of different words can be inserted that refer to station names, dates and times. In OVIS, the Speech Generation Module (SGM) is preceded by a Language Generation Module (LGM), which specifies the texts to be spoken. On the basis of syntactic and semantic information and the context of what has been previously mentioned, it computes the placement of accents and phrase boundaries. The naturalness of the output is increased by deaccenting given information and dividing sentences into smaller prosodic phrases.

1. Openbaar Vervoer Informatie Systeem

This paper describes a formal evaluation experiment of three speech output techniques, viz. conventional phrase concatenation, IPO phrase concatenation and diphone synthesis. Natural speech is used as a reference condition. The goal of this experiment was to determine whether IPO phrase concatenation is significantly better than conventional concatenation and to find out whether diphone synthesis would be an acceptable alternative. The next section will discuss the three techniques in some more detail.

Speech output techniques

Conventional phrase concatenation

One technique that is used quite frequently in commercial spoken dialogue systems is phrase concatenation, where generated texts are analysed and all words and phrases required to make these texts audible are recorded and stored in one version. This requires a relatively small effort.

The general disadvantage of this type of technique is that its use is limited to applications with a medium-sized vocabulary with a relatively fixed content. Moreover, this type of technique is unable to produce variability in prosody. An additional problem is that most commercial producers of such systems do not spend a lot of time and effort into perfecting the resulting output with respect to differences in loudness, tempo and pitch.

IPO phrase concatenation

For OVIS, the conventional phrase concatenation technique was extended to be able to deal with the prosodic variability computed by the LGM. The technique, which we shall refer to as IPO phrase concatenation, is different from the conventional approach in that several prosodically distinct versions of otherwise identical words are recorded. Six different versions are required to cope with the combined effects of accentuation and position relative to the prosodic phrase boundary. These versions consist of accented words in sentence-medial, minor-phrase final and sentence-final position and their unaccented counterparts. The intonation contours required for these prosodic versions have been determined using the IPO grammar for Dutch intonation ('t Hart, Collier & Cohen, 1990). There are at least two reasons why the IPO approach is likely to outperform conventional concatenation. Firstly, Terken & Nootboom (1987) have reported that correct accentuation facilitates comprehension. Their conclusion was that accenting given information confuses the listener. Secondly, Sanderman & Collier (1997) have shown that comprehension is facilitated by good phrasing.

Having learned from the imperfections of conventional concatenation, we have controlled the recordings very strictly to reduce variability in tempo, loudness and pitch. All units (words and phrases) have been embedded in dummy carrier sentences that resemble the sentences in the application domain. After recording, the material has been checked and altered when necessary, thus ensuring that no mismatches will occur after concatenation. A more elaborate discussion of IPO's phrase concatenation technique can be found in Klabbers (1997) and Theune et al. (1998).

Although this method provides more flexibility than conventional phrase concatenation, its use is also restricted to applications with a medium-sized stable vocabulary.

IPO diphone synthesis

Diphone synthesis is provided via SPENGI (IPO's SPEech synthesis ENGIne). It uses a synthesis technique called phase synthesis, a technique developed at IPO, which reduces the audible deterioration of PSOLA manipulations on the concatenated units (Gigi & Vogten, 1997). It uses a special strategy to determine the relative contribution of periodic and noise components of the synthetic signal, based on a very accurate pitch synchronous analysis of the amplitude and phase of the harmonic components of the input signal.

In two blind tests concerning the subjective evaluation of synthesis systems under telephone conditions, SPENGI was judged favourably on several aspects, including general quality, intelligibility and voice pleasantness, compared to other commercially available synthesis systems for Dutch (Rietveld et al. 1997, Sluijter et al. 1998).

The general opinion about speech synthesis is that intelligibility of current systems is quite good, but naturalness still leaves a great deal to be desired. In this experiment diphone synthesis is included to see how well it compares to phrase concatenation techniques, that do not require any signal manipulation. A great advantage of using diphone synthesis in commercial applications is that it is much more flexible in generating the output and it doesn't require the output to be specified beforehand.

Evaluation method

The experiment consisted of three parts: an intelligibility test, a fluency test and a final test about the general quality and suitability to the application. In the final part, the natural speech condition was excluded, since it is not a realistic option for actual use in applications like OVIS.

Subjects

Twenty subjects participated in the experiment. They were students from the Eindhoven Technical University, with no prior knowledge of speech technology.

Materials

All messages, twentythree in total, presented to the subjects consisted of train connections. There were three example messages which were used at the start of the experiment

and in the final part of the experiment, which had the following form:

ik heb de volgende verbinding gevonden ///
met de sneltrein / vertrek vanuit enkhuzen / om elf uur eenenvijftig /
aankomst in oosterbeek / om drieëntwintig uur twee ///
daar verder met de stoptrein / vertrek om dertien uur zeventien /
aankomst in stavoren / om drie uur achtentwintig ///
wilt u nog een andere verbinding weten ? ///

*I found the following connection ///
with the express train / departing from enkhuzen / at 11:51 /
arrival in oosterbeek / at 23:02 ///
there continue with the slow train / departing at 13:17 /
arrival in stavoren / at 03:28 ///
would you like to have a different connection ? ///*

Twenty shorter messages were used for the intelligibility and fluency tests, which were of the form:

met de sneltrein / vertrek vanuit vierlingsbeek / om negen uur zevenenvijftig /
aankomst in koudum molkwerum / om drieëntwintig uur dertig ///

*with the express / departing from vierlingsbeek / at 09:57 /
arriving in koudum molkwerum / at 23:30 ///*

The station names were balanced with respect to familiarity (correlated with the number of inhabitants of the city) and number of syllables. Some station names were highly confusable (*Heiloo* vs. *Heino* and *Oss* vs. *Olst*).

The fragments of conventional phrase concatenation were obtained from a commercial train timetable information system. Since these fragments were only available in telephone bandwidth, the fragments from the other speech output conditions were filtered to reduce their bandwidth from 8000 Hz to 3400 Hz. The IPO phrase concatenation, diphone synthesis and natural speech fragments all came from the same semi-professional female speaker, whereas the conventional phrase concatenation used a different female speaker.

Procedure

First, the subjects were presented with the three example fragments, to get a general indication of the different speech output conditions. The natural speech output condition was not included in these examples.

The experiment continued with the intelligibility test. All subjects listened to five fragments of each speech output condition, so twenty fragments in total. The order of the speech output conditions was balanced over all subjects, so that one speech condition occurred five times in the first block, five times in the second block, etc. This was done to avoid order effects. The fragments were also divided differently over the speech conditions for each subject, so that each station name occurred equally frequently in each speech condition. The task for the subjects was to write down the two station names that occurred in each fragment (The total number of observations n is 200, i.e. 5 fragments \times 2 station names \times 20 subjects). Due to the format in which the train connections were pre-

sented to the subjects, it proved impossible to ask them to write down the times as well, let alone the entire sentence. This touches upon a problem of formulation that is very important in commercial applications, but which will not be discussed here. After listening to five fragments of one speech output condition, subjects had to rate on a 7-point scale how they judged the overall intelligibility of the system. Subjects were instructed to take into account whether some words could not be understood, or whether it took a lot of effort to understand the message (n = 20, i.e., 1 speech output condition x 20 subjects).

In the fluency test, the subjects had to rate the fluency of each fragment on a 7-point scale. The fragments were the same 20 fragments that were used in the intelligibility test, but the speech output conditions were presented in a different order, and the fragments were again differently divided over the speech output conditions. We did not ask for naturalness, as this term is a subjective measure, encompassing many aspects of the speech, including voice pleasantness, friendliness, etc. Instead, we asked subject to rate the fluency of the fragments, which is a better term that focuses on the speech itself. The subjects were told they could take into account whether the speech had a faltering speaking style, or contained audible jumps in pitch (n = 100, i.e., 5 fragments x 20 subjects).

In the final part of the experiment, the subjects were again presented with the example messages. Per message they had to rate two questions on a 7-point scale: one concerning the general quality of the speech (n = 20) and the other concerning its suitability for the given application (n = 20). Here, subjects had to imagine actually calling such a commercial application. How much would they appreciate the speech output of the system in that case?

Results

In the discussion of the results, the speech output techniques will be referred to by codes. N stands for the natural speech condition, IC stands for IPO phrase concatenation, CC for conventional phrase concatenation and DS for diphone synthesis. In the first part of the intelligibility test, where the subjects had to transcribe 20 station names (5 in each speech output mode) some errors in the transcription occurred. Table 1 lists the percentage of items that were transcribed correctly.

Speech output techniques	% items correct (n = 200)
N	93
IC	95
CC	89
DS	75

A two-tailed t-test shows these differences to be significant. Table 2 displays the mean values for the each of the four subjective scaling questions and each of the four speech output conditions. In Figure 1 these are displayed graphically together with the standard deviations to make the difference more visible. The most important finding is

that the difference between natural speech and IPO phrase concatenation is indeed small in terms of intelligibility and fluency. The difference in intelligibility is not even significant ($F(1, 19) = 0.11, p > 0,05$), but the difference in fluency is, although the F-value is small ($F(1, 19) = 19.52, p < 0,05$). No comparison between these two output conditions was made with respect to general quality and suitability for the application, as natural speech is not a realistic option for use in a spoken dialogue system. IPO phrase concatenation clearly is the best alternative.

A second conclusion is that conventional phrase concatenation performs significantly worse than our more sophisticated technique, on all dimensions. Interestingly, the difference between conventional phrase concatenation and diphone synthesis is not as large as expected. With respect to fluency, the difference between these two techniques is not even significant ($F(1,19) = 1.45, p > 0,05$). Diphone synthesis performs worse on all dimensions.

	# observations	N	IC	CC	DS
Intelligibility	20	5.95	5.85	4.80	2.75
Fluency	100	6.16	5.41	3.19	2.72
General quality	20	n/a	6.05	3.90	2.25
Suitability for application	20	n/a	6.60	4.10	2.45

Table 1: Average subjective scores for all speech output conditions

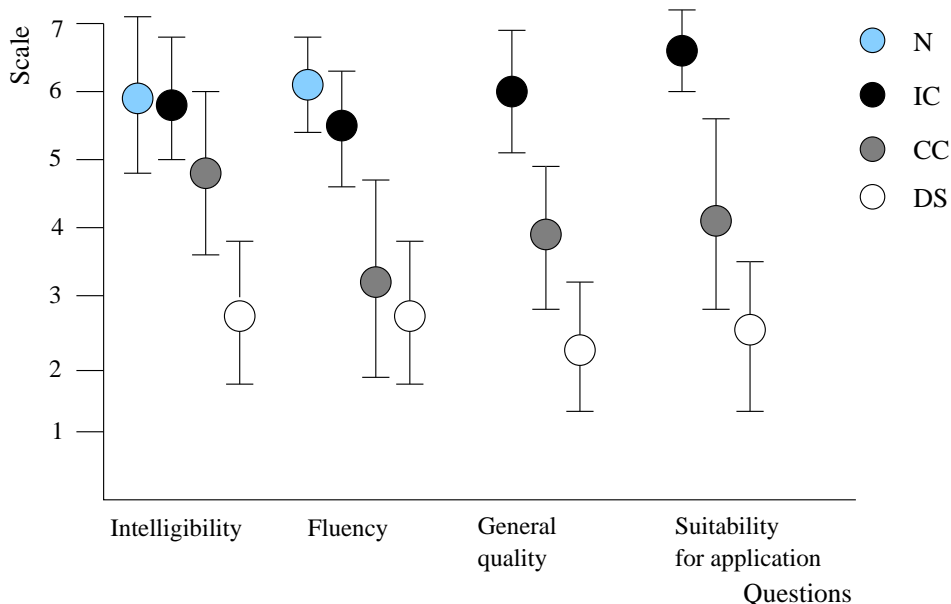


Figure 1: Average scores and standard deviations in the subjective evaluation

Discussion

This experiment has given quantitative evidence that the quality of IPO phrase concatenation approaches that of natural speech. It clearly provides a better output quality than conventional concatenation techniques. Naive listeners have a clear preference for the more sophisticated output. The extra effort needed to construct a prosodically sophisticated phrase database, is definitely worthwhile. Although conventional phrase concatenation is quite intelligible, it scores considerably less on fluency and general quality, which in our opinion is caused by a number of factors. One problem is the difference in tempo, loudness and pitch that occurs, and that in most commercial systems is often disguised by inserting longer pauses between the units. But more importantly, conventional concatenation can confuse the listener, as it is unable to deaccent given information and does not have the variability in pitch patterns required to phrase longer fragments properly. Both problems can be attributed to the fact that developers of commercial applications spend less effort on developing the speech output.

Diphone synthesis still has an inferior quality compared to the other techniques, on all dimensions. However, if the intended application requires more variability than can be handled with phrase concatenation, diphone synthesis is the only synthesis technique applicable. Therefore, one should continue to work on improvement of this technique so that its use in commercial applications may become more widely accepted. Currently, we are studying the occurrence of audible discontinuities at diphone boundaries (Klabbers & Veldhuis, 1998). The outcome of this study is that these discontinuities are caused by the spectral context and that these can be detected by using a spectral distance measure. In the future we intend to use a suitable spectral distance measure, which performed best in our experiments to cluster contexts and thus add context-sensitive diphones to the inventory to reduce the occurrence of discontinuities.

A second study that is under way is the development of a new duration module for Dutch that will compute the duration of the speech sounds in a more sophisticated way, taking into account higher-level factors such as accentuation, position in the word and position in the sentence.

Acknowledgements

This project is part of the Priority Programme Language and Speech Technology (TST), sponsored by NWO (the Netherlands Organisation for Scientific Research). Gerard Hollemans and Mili Docampo Rama are thanked for their statistical support.

References

- Gigi, E. & Vogten, L. (1997). A Mixed-Excitation Vocoder based on Exact Analysis of Harmonic Components. *IPO Annual Progress Report*, Eindhoven, Volume 32, 105-110.
- 't Hart, H., Collier, R. & Cohen, J. (1990). *A Perceptual Study of Intonation: An experimental-phonetic approach to speech melody*. Cambridge University Press, Cambridge.
- Klabbers, E. (1997). High-quality speech output generation through advanced phrase concatenation. *Proceedings of the COST Workshop on Speech Technology in the Public Telephone Network: Where are we today?*, Rhodes, Greece, 85-88.
- Klabbers, E. & Veldhuis, R. (1998). On the reduction of concatenation artefacts in diphone synthesis. *Pro-*

- ceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia.
- Rietveld, T., Kerkhoff, J., Emons, M., Meijer, E.J., Sanderman, A., Sluijter, A. (1997). Evaluation of speech synthesis systems for Dutch in telecommunications in GSM and PSTN networks. *Proceedings of EUROSPEECH'97*, 577-580.
- Sanderman, A. & Collier R. (1997) Prosodic phrasing and comprehension. *Language and Speech* 40(4), 391-409.
- 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.
- Theune, M., Klabbers, E., Odijk, J., de Pijper, J.R. & Kraemer, E. (1998). From Data to Speech: A Generic Approach. *Submitted to Natural Language Engineering*.

Appendix

heino	heiloo (4x), heijnen, hieno, hengelo
vierlingsbeek	veerlingsbeek, wierlingsbeek
(koudum) molkwerum	mookwerum, monkwering, monkwerum, mookwierum, morkweerdum, mollinkwerum, molksbierum, molkwierum, mankwerum,
maarn	marn, naarden
raalte	rapen, raten
(den haag) laan van nieuw oost indie	nieuw oost indie missing
zeist	zest
ermelo	wermelo
voorschoten	voorschotten, ...sloten
(vlissingen) soeburg	soedorf
(leeuwarden) kammingaburen	karlingaburen
warffum	wargum (2x), worvum, warfen
olst	onst, ost, omst
(hoogezand) sappemeer	velpermeer
lelystad	leliestand
nieuwerkerk (a/d ijssel)	nieuwkerk
vlaardingen oost	vlaardingen centrum

Table 2: Errors in transcription of station names