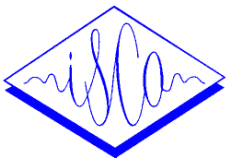

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

THE WAXHOLM APPLICATION DATABASE

Bertenstam J., Blomberg M., Carlson R., Elenius K., Granström B., Gustafson J., Hunnicutt S., Högberg J., Lindell R., Neovius L., Nord L., Serpa-Leitao de A. & Ström N. (Names in alphabetic order)

e-mail: waxholm@speech.kth.se
Dept. Speech Communication and Music Acoustics,
KTH, Box 70014, 10044 Stockholm, Sweden

ABSTRACT

This paper describes an application database collected in Wizard-of-Oz experiments in a spoken dialogue system, WAXHOLM. The system provides information on boat traffic in the Stockholm archipelago. The database consists of utterance-length speech files, their corresponding transcriptions, and log files of the dialogue sessions. In addition to the spontaneous dialogue speech, the material also comprise recordings of phonetically balanced reference sentences uttered by all 66 subjects. In the paper the recording procedure is described as well as some characteristics of the speech data and the dialogue.

INTRODUCTION

WAXHOLM is a demonstrator spoken dialogue system in which we apply our research on speech recognition and speech synthesis. The system uses visual and auditory means to provide information on boat traffic, accomodation, etc. in the Stockholm archipelago [1], [2], [3]. The application has great similarities to the ATIS domain within the ARPA community, the Voyager system from MIT [4], and similar tasks in Europe, see [5] for an overview.

Data has been collected using the system with a Wizard-of-Oz replacing the speech recognition module. The collected database consists of utterance-sized speech files that are stored together with the text entered by the wizard and the corresponding phonetic labels. A complete log of the dialogue session is also stored. The acoustic-phonetic data-base also consists of phonetically rich reference sentences uttered by all subjects. Thus, the data collected provide a good basis for studies of spontaneous speech and related phenomena, as well as read speech material suitable for inter-speaker comparisons, speaker adaptation experiments, etc.

This paper describes the experimental set-up and the characteristics of the speech and dialogue data collected so far.

WIZARD-OF-OZ EXPERIMENTS

In the data collection the speech recognition has been replaced by a Wizard-of-Oz (Fig. 1). This section describes the experimental set-up and procedure.

The subjects were seated in an anechoic room in front of a display with a uni-directional cardioid electret condenser microphone mounted on top of it. The wizard was seated in an adjacent room facing two screens, one displaying what was shown to the subject and the other displaying system information. The subjects were all aware of the fact that the wizard replaced the speech recognition. Each utterance was digitally recorded at 16 kHz and stored together with its respective text and, later, label file. All system information was logged during the data collection sessions making it possible to replay the dialogue including the graphics.

An experimental session started with a system introduction presented in text on the screen. The text was also read by speech synthesis, thus permitting the subject to adapt to the synthetic voice. The subject practiced the push-to-talk procedure reading a sound calibration sentence and a few test sentences followed by eight phonetically rich reference sentences. The reference sentences were designed to cover both common and uncommon phonemes. However, /ʏ/ and retroflex /ɭ/ are not represented. Each subject was provided with three information retrieval scenarios. Fourteen different scenarios were used altogether. The first scenario, which was the same for all subjects, is presented below.

> **Scenario 1**
> It's a beautiful summer day and you are in Stockholm. You decide that you'd like to go to Vaxholm[1].
> Your task is to find out when the boats leave for Vaxholm this evening.

The scenarios were presented to the subject both as text on the screen and with speech synthesis. The subject was instructed that the scenarios are starting points for further exploration of the system capabilities. It was not emphasised that the subject should regard the scenario as a task to be completed as fast as possible, encouraging the subject to use the system beyond the scope of the scenario. Utterances resulting in extremely low or high sound levels were automatically detected by the system, which prompted the subject for a repetition.

After the experimental session, the subject filled in a questionnaire with questions about age, weight, height, profession, dialect, speaking habits, native tongue, comments about the experiment, etc.

So far, some 1900 dialogue utterances have been recorded containing 9200 words. The total recording time amounts to 2 hours and 16 minutes, one third of which is labelled as pause. One fourth of the recording time pertains to the calibration and reference sentences.

[1] "Vaxholm" is a town in the archipelago. The spelling "Waxholm" is used in the name of a boating company and for some boats in their fleet.

RESULTS

Speaker characteristics

Initially, 66 different subjects, of which 17 are female, have participated in the data collection. The majority of the subjects, 43, were 20-29 years old while 4 were 30-39, 10 were 40-49 and 9 were more than 50 years old.

Most subjects are department staff or undergraduate students from the School of Electrical Engineering and Information Technology, KTH.

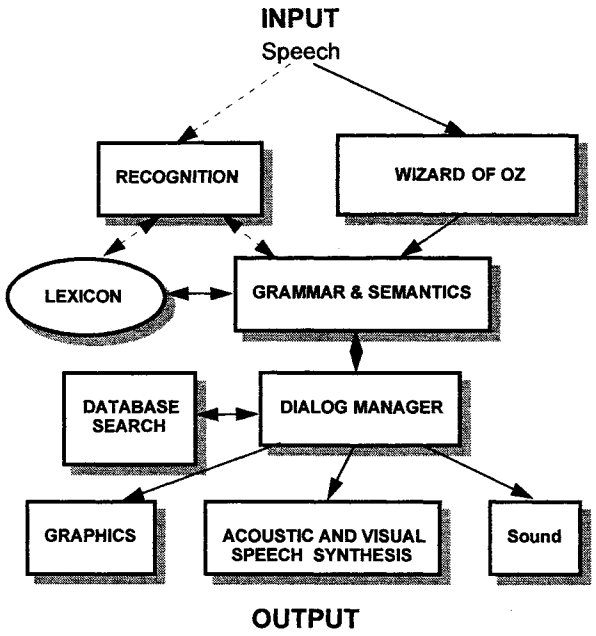

Fig. 1. The modules of the WAXHOLM spoken dialogue system.

The subjects have given a written description of their dialects or accents. Many speakers do not seem to appreciate the full value of their deviations from standard Swedish. That is, although some subjects have a discernible non-standard Swedish dialect they did not indicate it in the questionnaire.

Several dialects are represented but at least 50% of the speakers of standard Swedish have a Stockholm area dialect. A few foreign accents are also represented: Portuguese, Finnish, Norwegian, and American English.

Speaking styles and emotions

There is a noticable spread in speaking styles. The introductory test and calibration sentences sound very different from the following dialogue for some speakers. Many subjects alter speaking style when switching from read to spontaneous speech.

Most phonemes have a longer duration in the dialogue speech. Especially the phonemically long vowels tend to be prolonged in the spontaneous speech. The main reason for this may be that the subjects were slowed down since they were solving a task. Another sign of this is that the mean duration of sentence internal pauses were longer for the dialogue speech than for the read speech. A further reason may be that some speakers, although they knew they were talking to a Wizard spoke slowly and elaborately, perhaps because the system had not understood sentences that seemed very simple to the subject. Another factor for the elaborate speech might be the rather unnatural set-up, with the subjects alone in a soundproof booth with a microphone and a terminal.

Intra-subject variations in tempo and loudness are sometimes quite considerable. Extralinguistic phenomena, such as hesitations, pauses, coughs, and sounds of breathing, are also part of the speaking style.

The dialogue contains spontaneous reactions reflected in the subjects choice of words, the occurrence of extralinguistic sounds and emotional cues in the voice character. As many as five of the subjects have what could be considered "laughter in their voice." Explicit and muffled laughter occur in a few cases. Anger can be heard in some recordings and also more subtle indications of irritation or aggressiveness, even though the subject does not always express his or her emotional state in words.

Voice characteristics

A speech pathologist and a phonetician listened to the recorded speech and made a voice evaluation of the 66 speakers. As many as 33 speakers have a creaky voice character, especially in phrase endings, while eleven also have some grating. The voicing onset is hard in 27 of the voices.

In five of the voices, the voice register is unstable and a voice break occurs in two. The general term "hoarse" was used for ten of the voices. Twelve of the voices are considered to be carrying very little energy, while five are considered to be strained. These aspects are not pathological deviances, but can influence the performance of a speech recognition system.

Some of the young male speakers use a very low fundamental frequency at the bottom of their registers. Examples of rough, tense, weak, and screaming voice characters are also present in the speech material. Only one speaker sounded as if he had a cold.

Word frequencies

The word frequency ranking of the WAXHOLM corpus is naturally different from that of the KTH corpus (150 million Swedish words) which consists mostly of newspaper text, but also of text from novels, educational books and almost 5 million words from speeches in the Swedish parliament. The total number of unique words is 1.88 million compared to the about 600 different words in the WAXHOLM dialogue corpus.

The top ranking WAXHOLM words almost make up the sentence "jag vill åka till Vaxholm" (I want to go to Vaxholm), which indicates the influence of the domain and the scenarios given to the subjects. Though many of the most frequent words in the large corpus have a high rank in the dialogue corpus, the most frequent one, *'och'* (and), only has rank 35. It is also interesting to note that although there are only 600 unique words in the dialogues, 9 of them cannot be found in the large KTH corpus. Five of these are names of small ports. Of the other four, *'vartifrån'*, is a spoken language variant of *'varifrån'* (where from) while the others may be seen as very typical for the WAXHOLM domain.

The 10 most frequent words cover 35% of all words in the dialogues and the 200 most frequent words cover 92%.

Labelling and phonetic transcriptions

The recorded utterances are labelled on the word, phoneme, and phone levels with links between the levels. In this way, it is easy to extract the phonetic realisation of the words, as well as the word affiliation of individual phones or phonemes. We use an automatic labelling and alignment procedure, described by Blomberg & Carlson [6]. In that method, a lexicon and a set of rules designed for text-to-speech applications [7] are used for the generation of a base form phoneme transcription of an utterance. Optional word pronunciations are added and optional phonological rules are applied. The rules have proven to be especially important at word boundaries. The estimated phonetic transcription of a particular utterance is obtained as a result of the alignment procedure. The output of the automatic alignment procedure is manually corrected.

The phoneme alphabet is essentially identical to that used in the KTH text-to-speech system [7]. Some extensions have been added to the text-to-speech inventory in order to account for extralinguistic and non-speech sounds. Another difference is that plosives have been split into an occlusion and a release.

Certain words have been pronounced in many different ways: varying tonal accents, consonant deletions, relaxed pronunciation forms, word boundary coarticulations, etc., give rise to a large number of unique transcriptions of a word. Also non-canonical pronunciations have sometimes been used, especially for names. A common word in this application, *'skärgården' (the archipelago)*, has been transcribed in 25 different ways. Function words are often quite reduced. Vowels, consonants, and occasionally even syllables, can be deleted in these words. The initial and final phonemes of the words are often modified due to coarticulation with the previous and the following words.

The domain specific word frequency distribution influences the corresponding phoneme occurrence frequencies. The reference sentences are, as previously mentioned, designed to contain uncommon phonemes. The use of these sentences during training of a phoneme library for recognition will therefore raise the coverage of low-frequency phonemes. One remaining problem is the very limited variability of phonemic context of these phonemes due to the low number of different sentences. This will change the acoustic characteristics from neutral positions towards positions given by the surrounding phonemes. Further, the fact that these phonemes pertain to read rather than spontaneous speech has implications for their spectral properties.

Extralinguistic sounds

Extralinguistic sounds are transcribed manually during the post-processing of the data. The extralinguistic categories that are considered are interrupted words, audible inhalations and exhalations, clicks, laughter, lip smacks, hesitations and hawkings.

Inhalations, which often occur in combination with smacks, are the most common extralinguistic events. All but a few inhalations are utterance initial. There are also aspirated non-speech segments, labelled as parts of pauses, which are generally less prominent than the initial inhalations. Exhalations occur in utterance final positions and, to some extent, in mid-utterance positions. Most utterances end with a relaxation gesture, that is, with centralised and often aspirated segments. When the final aspiration is strong, it is labelled as an exhalation.

Inserted vowel sounds are also labelled. They occur when a consonant constriction is released. More than 80% of the inserted vowel segments are word-final, about 10% can be found in compounds at intra-word boundaries and quite often they are found in utterance final position. The phonetic context of an inserted vowel segment is often /r, l, m, n/. The inserted vowel segments are unevenly distributed over the subjects. As few as 10 speakers make more than half of all vowel insertions: 138 insertions out of 238. About one fourth of the speakers have no inserted vowels at all. Thus, vowel insertion is a speaker-specific feature occurring in clear speech.

Hesitations are commonly found in utterance- or sentence-initial positions. The major part of the remaining cases are found in conjunction with place names as, for example, in the utterance "I would like to go to *uhm* Vaxholm." This could either be a common dialogue phenomenon or an artefact due to the fact that the dialogue is not spontaneous.

Dialogue analysis

Subject performance

A total of 66 subjects participated in the experiment. Each subject was presented with 3 scenarios. A total of 198 scenarios were recorded and analysed. Each scenario required that the user solved one to four subtasks. A subtask could be that the subject had to request a timetable, a map or a list of facilities. Each subtask, in turn, required specification of several distinct constraints, such as departure port, destination port and departure day, before the subtask could be solved. The subjects had to provide the system with up to ten such constraints, with a mean of 4.3, in order to solve a complete scenario.

The database contains 265 subtasks and about 84% of these were solved by the subjects. In 75 percent of the cases, 199 out of 265, the subjects had completed a subtask after one to five utterances. The subjects needed about 7 utterances to solve one scenario. After the task was completed several subjects continued to ask questions in order to test the system. About 3 additional utterances were collected this way. In 42 cases a scenario could not be completely solved by a subject (21%). In half of these, 21 scenarios, some of the subtasks were solved by the subjects.

The average utterance length was 5.6 words. The average length of the first sentence in each scenario was 8.8 words. The utterance length distribution shows one maximum at two words and one at five words. One reason for this distribution is that many of the utterances were subject answers to system questions. As an example, one type of system question was "Which port

would you like to go to/from?". A typical answer to this question was "To/From Stockholm" or "I want to go to/from Stockholm." (The infinitive mark is left out in Swedish).

We can find a few examples of restarts in the database due to hesitations or mistakes on the semantic, grammatical or phonetic level. However, less than 3% of the utterances contain such disfluencies. Some of the restarts are exact repetitions of a word or a phrase. In some cases a preposition, a question word or a content word is changed. We also find repetitions of incorrectly pronounced words. About one fourth of the restarts occur in interrupted words, that is, in words that are not phonetically completed.

System response analysis

The Waxholm database contains approximately 1900 dialogue turns. After the first 37 sessions, the system went through a major revision. The first phase included approximately 1000 subject utterances. The system responses "I do not understand" and "You have to reformulate" occurred in 35.8 % of the system responses. In the second phase, the dialogue manager was updated as well as the scenarios. In this phase, 31 subjects produced 900 utterances. The improved system failed to understand 20.9% of the time, an improvement of 15%. It should be noted that this system response in some cases also is the correct one.

Most of the questions from the system occurred when the system predicted that the subject wanted a timetable displayed. In these cases, the distinct constraints were evaluated, and if some information was missing, the system took the initiative to ask for this information. The subjects answered the system questions in 95.4% of the cases. Thus, the subjects were quite co-operative and rarely, one percent, used the possibility to change the topic during the system-controlled dialogue. In a more realistic environment, using speech recognition as input, the system might misunderstand the user's goal, and topic changes by the subject will become more frequent.

The most serious problems occurred when the system failed to 'understand' an utterance from a subject. The first system response was a simple "I do not understand" utterance. If the system failed to understand once more, the system elaborated more on the problem. First, the subject was informed where it failed to understand, if it was a linguistic problem. Second, the system asked the user to use a complete sentence next time. The following utterance from the subject was used to evaluate whether the system-predicted topic actually agreed with this new utterance or whether the topic should be changed. The system responded 'I don't understand' 575 times corresponding to 268 occasions where consecutive repetitions are counted as one occasion. In 50% of the cases the system recovered after one additional utterance.

FINAL REMARKS

The analysis of the dialogue material will be valuable in the development of intelligent system response generation. The spontaneous speech data collected provide good training material for the speech recognition module, which will be tested and evaluated within the framework of the WAXHOLM application.

The intended use of the test sentences is to perform experiments with speaker characterisation and fast speaker adaptation, in which phonetically rich speech data is desired to obtain high performance. The test sentences have been used in a study of human speaker recognition [8]. Moreover, the speech data is used for acoustic-phonetic studies of context and position-dependent spectral phoneme variations.

We have updated the system and started a new data collection phase using the speech recognition module instead of a Wizard. The users encounter a more realistic system set-up featuring a graphical interface visualising the domain of the application and the system capabilities [2]. Thus, the self-explanatory interface permits non scenario driven data collection.

ACKNOWLEDGEMENTS

This work has been supported by grants from The Swedish National Language Technology Program.

REFERENCES

[1] Blomberg M., Carlson R., Elenius K., Granström B., Gustafson J., Hunnicutt S., Lindell R., Neovius L. (1993). "An experimental dialog system: WAXHOLM", Proc. of Eurospeech '93, pp 1867-1870, Berlin.

[2] Carlson R (1994). "Recent developments in the experimental "WAXHOLM" dialog system." Proc ARPA Human Language Technology Workshop, pp 207-212, Princetown, New Jersey.

[3] Bertenstam J, Beskow J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, Nord L, de Serpa-Leitao A & Ström N. (1995). "The Waxholm system - a progress report", Proc. ESCA Workshop on Spoken Dialogue Systems - Theories and applications, pp 81-84, Vigsø, Denmark.

[4] Glass J, Flammia G, Goodine D, Phillips M, Polifroni J, Sakai S, Seneff S & Zue V. "Multilingual spoken-language understanding in the MIT Voyager System.", Speech Communication. (To be published.)

[5] Proceedings from the ESCA Workshop on Spoken Dialogue Systems - Theories and applications, May 30-June 2, 1995, Vigsø, Denmark.

[6] Blomberg M., Carlson R. (1993). "Automatic labelling of speech given its text representation.", Papers from the Seventh Swedish Phonetics Conference, RUUL, #23, pp 61-64, Uppsala.

[7] Carlson R., Granström B. & Hunnicutt S. (1991). "Multilingual text-to-speech development and applications.", Ainsworth AW, ed, *Advances in speech, hearing and language processing*, London: JAI Press, UK.

[8] Carlson R. & Granström B. (1994). "An interactive technique for matching speaker identity." Papers from the Eighth Swedish Phonetics Conference, Working Papers 43, Dept. of Linguistics, University of Lund, pp 41-45.