Survey of spontaneous speech phenomena in a multimodal dialogue system and some implications for ASR

Louis ten Bosch , Lou Boves

Radboud University Nijmegen, The Netherlands
{l.tenbosch, l.boves}@let.kun.nl

Abstract

Audio recordings of speakers using speech-driven systems show phenomena that are characteristic for on-line speech responses, such as out-of-task utterances, self-talk and speech disfluencies. This paper focuses on a survey of these phenomena as they were recorded during interactions by subjects using a multimodal system, and reports on experiments concerning the treatment of these phenomena for automatic speech recognition. This study is a starting point for the study of a richer set of on-line phenomena in speech addressed to multimodal systems and the implications for automatic speech recognition.

1. Introduction

This paper addresses spontaneous speech phenomena observed in recordings that were made from users who interacted with a multimodal system for an architectural design application. We have annotated the database with labels that distinguish several different types of phenomena. The first type comprises disfluencies (restarts, repeats, hesitations, filled pauses) which are due to the on-line character of interactions. The second type of phenomena comprises various 'noise' events (such as loud breath sounds, lip smacks, microphone clicks, background speech/noise). The last type concerns non-grammatical and out-of-task responses (such as self-talk).

We have started annotating spontaneous speech phenomena to investigate their distribution in a real-life recorded database, and to investigate their impact on the performance of ASR in a multimodal interaction system. In this paper we provide a first distribution, and we investigate how ASR performance can be improved by including a garbage model to handle disfluencies and other spontaneous speech phenomena. An additional motivation to specifically annotate the different types of phenomena is that some of these (in combination with prosody) may help to detect and diagnose problems in human-machine communication [2, 9]. We intend to use the (continuously growing) database introduced in this paper for developing procedures to use disfluencies and prosodic information as additional information for a dialogue management module.

The speech analysed in this paper was addressed to a multimodal system while the users had to perform a specific design task. The experiments were performed in the context of the European IST project COMIC[1]. The COMIC project aims at the study of natural and intuitive interaction between humans and multimodal systems. A system is envisaged that reacts in an intuitive and plausible way to the intentions of a user, expressed in any combination of available input channels. In order to move beyond the conventional map and form-filling applications, an architectural design application is implemented, which is instantiated in the form of a bathroom design task. More specifically, a multimodal system has been designed to help a non-expert, cooperative user to enter the relevant details of the desired bathroom using pen and speech for input. During the first phase of the interaction, the user is asked to draw the ground plan, i.e. the walls, windows and doors and to specify the lengths, heights, and widths of these objects. Also the direction in which the door(s) open(s) must be specified. This ground plan fully specifies the constraints for placing sanitary ware and additional bathroom furniture. In the second phase of the interaction, the user can discuss and compare a number of design options, available via guided browsing, to complete the bathroom ([6, 8, 10]).

The multimodal system in this study consisted of several modules: two input recognisers (a speech recognition module and a recogniser for pen input), a fusion module that does natural language processing and merges the information from the recognisers, a dialogue-action management module, and an output module, rendering beautified graphical information and speech prompts asking for additional information or explaining what the user can do next. A strict turn-taking protocol was implemented to ensure that the recognisers and the fusion module were able to process the multimodal input of the user in a coherent way. This protocol takes the moment at which the acoustic system prompt finishes as a starting point for the user's turn. After the prompt, the user must start providing input (either speech or pen or both) within a time window of two seconds. If the user starts within this two-second window, a window of five seconds is given to complete the response. Barge-in was disabled.

In the current implementation of the system, the interaction is based on a system-driven dialogue strategy. System-driven dialogues were chosen because inexperienced users need help for completing a complex task that they do not know how to perform. Moreover, this strategy can be used to gently push users towards expressions within the range of what the speech and pen input recognisers can handle.

In [8] a summary is presented of the performance of the input recognition modules in the system. Here we go into more detail on disfluencies and other phenomena in the speech that affect ASR. In the last decade, considerable attention has been paid to the phonetic and prosodic properties of disfluencies (e.g. [5, 7]). Although these studies provide a description of disfluencies in spontaneous speech, the results cannot be used directly to build acoustic models in ASR.

In the following sections, we will first provide an overview of the phenomena that were encountered in the audio recordings made, and introduce the ASR. Next, experiments are described that were performed to adapt the

[1] www.hcrc.ed.ac.uk/comic

ASR to better cope with the various speech phenomena. A discussion and conclusion are presented in the final section.

2. Description of the data

The speech data used in this study are speech files recorded in interactive sessions by 28 native, German speaking subjects. In a human-factors experiment [8], subjects had to enter the blueprint of three different bathrooms in one single session. The recordings from this ongoing experiment were converted to a growing database comprising audio files, verbatim transcriptions and annotations of the non-verbal audio events, such as non-speech sounds produced by the speaker, noise, back-channels, loud breath sounds, truncated words, repeats, self-talk, out-of-grammar speech, or background speech from other speakers. Back-channels such as 'hmm', which in human-human communication serve as a signal to the speaker that the listener is paying attention and understands the speaker, are considered as a special type of utterance, different from filled pauses. This collection of experimental data will here be referred to as the EXP-database. The total number of logged wave files (and annotation files) in the EXP-database is 2728.

Apart from the data recorded during the formal HF experiment, additional recordings have been made during the development of the multimodal system under the same conditions as the eventual experiment. This extra set is referred to as the ADD-database. This database contains 563 wave files. The total number of utterances in the union of both databases is presently 3291.

Each utterance is stored as a header-less, 16 bit/sample, 16 kHz mono sampled data file. The length of the speech files was determined by the automatic end pointing of the ASR system. Each audio file contains the recording between the moment of opening and closing of the microphone. The microphone was opened by command of the dialogue-action management module in the system, while the microphone was closed by the end pointing in the ASR module. The average duration of the wave files is 3.55 seconds.

Not all of the audio files contain speech, for two reasons. First, in about half of the system prompts the user was asked to provide information for which the pen is the obvious input channel. For example, at some point in the interaction the system instructs users to draw walls, an activity that obviously requires pen input. Most users do not speak during this drawing activity – some however, produce backchannel speech or self-talk (e.g., '[hmm] the wall goes here', 'maybe I had better start [hmm] at the other side'). Second, although lengths and other sizes could be entered by speech and pen, most subjects ended up using the modality with the lowest error rates (which turned out to be the pen modality, cf. [8]).

Table I shows the distribution of the number of lexical items per utterance (transcription length) for the databases EXP and ADD taken together. From the total number of 3291 wave files, 1419 contain just silence (the row with Transcription Length 0 Table I). Of the remaining 1872 wave files, the majority (86 %) of the utterances comprise just one (often 'yes', 'no'), two (such as in 'zwei meter', 'ja richtig') or three words (mostly a length, e.g. 'drei meter zehn'). In this respect, the utterances are similar to what has been observed in spoken dialogue systems.

Of these 1872 utterances that contain at least one word, 812 (43%) contain either a mix of speech and non-speech sounds, or speech that was not directly addressed to the system, but was nevertheless processed by the ASR. These 812 'noisy' utterances can be categorised according to the properties shown in Table II. In case an utterance matched multiple properties, the most salient property was chosen for its categorisation.

From Table II it can be deduced that Restarts occur rarely in the combined EXP+ADD database. In fact, we encountered only 33 cases. Examples are (square brackets denoting the reparandum): '[cen…] centimeter', '[hun…] hundert centimeter', and '[zwei…] zweihundert centimeter'. Repeats (of words of entire phrases) occur 37 times in the database. Examples are 'nein nein nein' and variants (15 occurrences), 'ein meter ein meter' and other variants of lengths (8 times), 'ich möchte ich möchte das löschen [...]', 'ja ja', '[...] ja [mmm] ja'. Almost all repetitions express frustration and irritation of a subject, caused by repeated recognition errors.

Table I. Distribution of utterance transcription length (first column) in the combined database EXP + ADD.

Transcription Length	# Occurrence	Proportion
0	1419	0.43
1	1049	0.32
2	365	0.11
3	210	0.06
4	154	0.05
5	65	0.02
6	22	0.01
7	6	0.00
8	1	0.00

Table II. Distribution of the 812 'noisy' utterances (utterances with non-speech sounds or with speech not addressed to the system) as a function of the type of phenomenon.

Number of occurrences (total 812)	Description of category
263	Utterances with single audible breath only.
143	Utterances with at least one filled pause ('hmm').
126	Utterances with audible breath sounds before, within or after the utterance.
88	Self-talk, out-of-task speech, mostly out-of-grammar.
70	Utterances with truncation/restarts/repeats of words.
59	Recordings with back-channels only.
45	Recordings with non-speaker noise (including background speech) only.
11	Recording with clicks only.
7	Recording with lip smacks only.

Many inexperienced speakers slow down the speaking rate after a recognition error of the ASR module – probably hoping that the ASR performance will increase. In the EXP database, we found two speakers (out of 28) exaggerating this strategy by explicitly pausing within words ('centi [pause] meter'). We found 31 clear examples of such within-word pauses. In the majority of cases (26), the duration of these within-word pauses exceeds 300 ms. Almost all speakers (25) consistently use some utterances with pauses *between* words that are longer than 300 ms seconds (e.g. 'vier [pause] meter').

Table III gives an overview of the distribution of non-speech phenomena in the EXP and ADD parts of the total database.

Table III. Distribution of utterance types over the two databases.

Database	EXP	ADD	EXP+ADD
total # recordings	2728	563	3291
# silent recordings	1217	202	1419
# utterances with speech only	832	228	1060
# utterance with speech & non-speech	679	133	812

3. The automatic speech recogniser

The automatic speech recogniser that was used in the multimodal system is based on HTK 3.1. In order to have HTK operating as a 'client' module in the system, the original HTK code was substantially modified to change the autonomous behaviour of HTK and to let HTK interact via various communication pools. The original 'on-line' HTK audio recording facility was kept to capture the audio input from a wire-connected close-talk head-mounted microphone.

The acoustic models used in the experiments were trained with the German SpeechDat database. For the training about 80,000 utterances (8 kHz sampling freq.) have been used. Models were based on 12 MFCC coefficients and one energy component, plus delta and delta-delta components. No cepstral or energy normalisation (utterance level or speaker level) was performed. In the tests reported here, gender-independent, context-independent models were used with 8 Gaussians/state, with 3 emitting states per phone. During the tests, the 16 kHz audio data were processed by the same filter bank as the one used during training. There was no on-line cepstral/energy normalisation available.

In order to handle non-speech events, two acoustic models were created to model 'general speech'. These models were also trained on SpeechDat. A garbage phone 'gp' has been modelled as a 3-state phone model; the garbage phone sequence 'gps' has been modelled by a 10-state HMM. The garbage models were bootstrapped by the schwa-model and by a sequence of schwa-models, respectively. Next, both models were trained in 3 full Baum-Welch re-estimation passes, after replacing randomly chosen single phone segments in the phone transcriptions by 'gp', and replacing randomly chosen phone sequences of length 3 by 'gps'. Afterwards, these newly trained general-speech models were added to the set of already trained phone models. In the lexicon, a garbage word **garb** was phonetically defined in terms of the 'phones' gp or gps (see below).

For the specific bathroom application domain, no language model was available. Therefore, the eventual LM that was used in the experiments was based on a handcrafted regular grammar (in BNF). This grammar contained 135 words and the grammar perplexity as calculated with HTK was 9.6. Since this grammar does not provide for optional inter-word garbage, we constructed a tool to extend the HTK search network with optional inter-word garbage entries. To that end, the 'clean' decoding lattice has been extended, by introducing one optional garbage word between each word pair. Parallel to each original arc **w1-w2** in the network, three new arcs **w1-garb**, **garb-garb**, and **garb-w2** were added to the lattice. In the present implementation all transitions *to* a garbage word were given the same garbage entrance penalty, which made for an additional tuning parameter in the experiments, next to the word entrance penalty.

4. Experimental results

In this section, we present recognition results based on the modelling and tuning of the two different garbage models. We will focus on string error rates, since string error rates are often of importance for evaluating the ASR-output in terms of the semantic interpretation on utterance level. (In these tests, a string accuracy of 65 % corresponds with a word accuracy of approximately 75-77 %.) The garbage penalties were tuned by optimising the word accuracy on the development set of the 361 (228+133) non-silent utterances in the ADD database. The 1511 non-silent utterances from the EXP-database were used as test set.

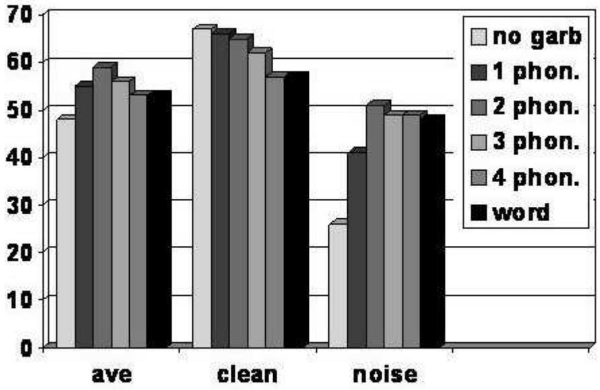

Figure 1. Results of the recognition experiments (string error rates) for all data (ave), the clean data only (clean) and the noisy data (noise) for different implementations of 'garbage'.

Fig. 1 shows string accuracies (SACC) for the evaluation set. The table presents accuracies (percentages) when a garbage model was absent, when the lexical entry **garb** was modelled by a sequence of 1, 2, 3, or 4 garbage phones (gp) or by a single 10-state model gps (indicated by *word*).

From the figure it is clear that there is a large difference in the results for 'clean' and 'noisy' utterances: adding garbage arcs deteriorates the performance for the utterances that are in the grammar. At the same time, it is obvious that utterances which do not fully conform to the grammar gain enormously from the addition of garbage arcs. In our test database the performance increase for noisy data outweighs the decrease for the clean data. The highest SACC was obtained in the case where the garbage word consisted of two garbage phones.

Table IV gives an overview of the ASR performance (SACC) on the entire EXP-database, but now as a function of the expectation (displayed column-wise) of the system prompt to which the utterance was a reply. All entries denote numbers of occurrence. Five expectations were used: Size, Confirmation (denoted by 'Conf'), Wall, Door, and Window. An utterance of category 'Size' is a reply to a system prompt asking for a size, and is therefore expected to contain information about a length (e.g. 'zwei', 'zwei meter', 'zwei meter dreizehn'). A confirmation refers to a 'yes/no' reply. The columns for Wall, Door and Window refer to utterances produced while users were drawing. Along the rows, the number of utterances is indicated for which the ASR was incorrect (inc.), correct (cor.), whether the utterance out-of-grammar (oog.), and whether there was just silence of any noise (sil/n), respectively. The number given for correct and incorrectly recognized responses refer to word strings. the multimodal character of the interaction The large number of silent audio recordings (690) when the system prompt asks for a 'size', are a to the multimodal character of the interaction.

Table IV. Overview of the ASR performance on the entire EXP-database, categorised according to the expectation of the system prompt. The entries represent numbers of occurrence. Numbers referring to incorrect ('inc') and correct ('cor') are string -based.

Exp	Size	Conf.	Wall	Door	Win-dow	Total
inc.	258	33	17	23	3	331
cor.	279	260	51	36	3	629
oog.	129	32	42	21	7	231
sil/n	690	53	462	207	125	1537
tot.	1356	378	569	287	138	2728

When the silent, noisy, and out-of-grammar inputs are left out of consideration, the recognition rate over all turns is 65.5 percent (string accuracy), in accordance with Figure 1, case 'clean data', '2 phones'.

5. Discussion and conclusion

A survey has been provided of the distribution of spontaneous speech phenomena in interactions of users with a multimodal system. The performance of the ASR is shown to be dependent on the structure of the garbage model. In the present experiment we used a unique garbage entry to deal with all phenomena. However, the ASR performance in on-line applications is expected to gain much more from specific, dedicated garbage modelling that is focussed on capturing the various acoustic phenomena.

The ASR performance (50-60 percent string error rate) that we obtained with what can be considered as off-the-shelf acoustic models is inappropriate in a realistic online application. Among ASR improvements that will be addressed in the near future are on-line acoustic normalisation, gender modelling, use of triphones, and the inclusion of special acoustic models for non-speech background noise. However, future research will specifically address the fact that disfluencies are not randomly distributed, but obey linguistic patterns [11]. We will investigate ways in which this knowledge can be exploited to select context-specific garbage modelling. One obvious way of doing this is to select the garbage model on the basis of the expectation that is generated by the dialogue-action manager. For example, when a user is prompted to draw something, one might want to use a model specialised for dealing with backchannels, while one would probably prefer a model for dealing with hesitations and filled pauses when the user was prompted to enter a length or similar information.

The results further raise the question how the modelling of speech disfluencies can be improved by using additional (e.g. paralinguistic) features. Disfluencies and utterance structure are associated with specific prosodic patterns (e.g., [3]). Since disfluencies, in combination with prosody, may provide useful information about annoyance or frustration [1], and help to flag problems in human-machine communication [2], future research will be focused on incorporating paralinguistic features in ASR for multimodal systems.

6. Acknowledgement

The research reported in this paper was supported by the IST project COMIC (IST-2001-32311).

7. References

[1] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Proc. ICSLP 2002*, p. 2037-2040.

[2] Batliner, A., Fischer, K., Huber, R., Spilker, J., and Noth, E. (2003). How to find trouble in communication. *Speech Communication*, vol. 40, pp. 117-143.

[3] Ferrer, L., Shriberg, E., and Stolcke A. (2003), A prosody-based approach to end-of-utterance detection that does not require speech recognition. *Proc. IEEE ICASSP*, Hong Kong, vol. 1, pp. 608-611.

[4] HTK. Url: htk.eng.cam.ac.uk.

[5] Nakatani, C. and J. Hirschberg (1994). A corpus-based study of repair cues in spontaneous speech, Journal of the Acoustical Society of America, Vol. 95, No. 3, pp. 1603-1616.

[6] Rossignol, S., ten Bosch, L., Vuurpijl, L., Neumann, A., Boves, L., den Os., E., de Ruiter, J.P. (2003). Human factor issues in multi-modal interaction in complex design tasks. HCI conference, Greece, June 2003.

[7] Shriberg, E., (1999). *Phonetic consequences of speech disfluency*. ICPhS-1999, San Francisco.

[8] Vuurpijl, L., ten Bosch, L., Rossignol, S., Neumann, A., Pfleger, N., Engel, R. (2004). Evaluation of multimodal dialog systems. *LREC workshop on Multimodal Corpora and evaluation.* Lisbon, 2004.

[9] Walker, M.A., Langkilde, I., Wright, J., Gorin, A., Litman, D. (2000). Learning how to predict problematic situations in a spoken dialogue system: Experiments with How may I help you? *Proc. NAACL-00*, pp. 210-217. Seattle.

[10] E. den Os and L. Boves. (2003) Towards Ambient Intelligence: Multimodal Computers that Understand Our Intentions. *Proceedings eChallenges 2003*, Bologna, 22-24 October 2003.

[11] Stolcke, A., and Shriberg, E. (1996). Statistical language modeling for speech disfluencies. In *Proceedings ICASSP*, vol. 1, pp. 405-408.