BMBF

Verbmobil

Verbundvorhaben

CONTRASTIVE ACCENTS – HOW TO GET THEM AND WHAT DO THEY LOOK LIKE

J. Haas, A. Kießling,
E. Nöth, H. Niemann,
A. Batliner

F.–A.–Universität Erlangen–Nürnberg

L.M.-Universität München

Vm

Report 89
Juli 1995

Juli 1995

J. Haas, A. Kießling,
E. Nöth, H. Niemann,
A. Batliner

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Institut für Deutsche Philologie
Ludwig–Maximilian Universität München
Schellingstr. 3
D–80799 München

Tel.: (09131) 85 - 7799
e-mail: kiessl@informatik.uni-erlangen.de

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 F/4 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

To be published in Proc. ICPhS'95:

CONTRASTIVE ACCENTS – HOW TO GET THEM AND WHAT DO THEY LOOK LIKE

J. Haas[1], *A. Kießling*[1], *E. Nöth*[1], *H. Niemann*[1], *A. Batliner*[2],
[1]*Lehrstuhl f. Mustererkennung (Inf. 5), Universität Erlangen-Nürnberg, Erlangen, FRG*
[2]*Institut f. Deutsche Philologie, L.M.-Universität München, München, FRG*

ABSTRACT

Automatic dialog systems tested with naive users are often confronted with special speaking styles, as e.g. words produced with emphatic or contrastive accent. Such utterances usually cause problems for word recognizers, because they were not included in the training data. It is thus important for the improvement of future systems to be able to collect utterances containing contrastive accents produced as natural as possible. We describe in this paper an automatic simulation system for provoking and collecting contrastive accents. With this system, 15 recording sessions were conducted; in total 205 word tokens produced either with default or with contrastive accent were collected. We discuss the results of an automatic classification as well as the relevance of extracted prosodic features for the marking of contrastive accents.

INTRODUCTION

While testing our automatic speech understanding and dialog system EVAR with naive users (via public telephone) the following situation was often observed: Because parts of the user utterance are not recognized correctly, the system delivers the wrong information. Usually, the user repeats the misrecognized words in a special, often excessive manner, using emphatic or contrastive accent. These utterances cause all the more recognition problems (not only for EVAR, but for all existing word recognition systems), because they were not included in the training data, and the dialog fails. Thus, there is a strong need for the collection of utterances produced with emphatic or contrastive accents and to take them into consideration during the training phase.

For the collection of words or phrases with contrastive accent it is essential that the data are produced as natural as possible. Asking speakers to read contrastive accents is a traditional [1] but suboptimal way. On the other hand, spontaneous speech corpora from human–human–dialogs contain very few contrastive accents. For example, in 20 investigated dialogs (approx. 60 min speech) of the VERBMOBIL–Corpus [4] no single contrastive accent could be observed. Another possibility for the collection of contrastive accents is to use the human–machine–dialogs conducted with the EVAR system. However, compared to all user utterances the occurrence of contrastive accents is not that high, and therefore very much effort had to be put on their identification.

In this paper we describe an automatic system with which a large amount of naturally produced contrastive accents can be provoked and collected. The system conducts dialogs with naive users by simulating an automatic speech understanding system in the domain of "train time table inquiries". It is designed to collect prosodic minimal pairs of words containing either the default word accent or a contrastive accent. In the second case, the position of the contrastive accent (either on the lexical word accent syllable or on a different one) can be induced by the system. It is thus possible to overcome the paradox to provoke spontaneously produced prosodic minimal pairs in an experimental environment.

THE SIMULATION SYSTEM

The simulation system is a Wizard–of–Oz–System where the role of the human wizard is played by the machine. Because it is no human wizard who can react on any possible user utterance in a flexi-

1.	System simulates correct recognition of the user utterance	
	(a)	System does not ask back (i.e. passing desired information) S: *"You can take the train at 10.47 ..."*
	(b)	System asks back S: *"You want to go to Hamburg?"*
2.	System simulates recognition error	
	(a)	System provokes contrastive accent on the word accent syllable S: *"Do you want to go to Hamburg or to Homburg?"*
	(b)	System provokes contrastive accent on the second syllable S: *"You want to go to Hamberg?"*
	(c)	System provokes a distinct (emphatic) pronunciation S: *"Where do you want to go?"*

Figure 1: *Possible system reactions following the first user query.*

ble manner but a simple computer program, the structure of each dialog conducted with the simulation system is heavily restricted. In any state of the dialog, the system has to react in such a way that there is no other possibility for the user than to behave in an expected, predefined manner. On the other hand, it is essential to prevent the user from realizing that he/she is *not* communicating with a 'normal' automatic system. One way of doing this is to produce a well-balanced proportion of 'artificial' recognition errors in the system's output. The speaking style of the users should not be influenced, and therefore, the output of the system is always presented in textual form on the screen; no synthesized speech is used. To prevent the users of becoming bored too soon and to get as natural utterances as possible it is important to provide them with a good amount of different alternating system reactions as well as to grant them from time to time a sense of achievement by passing the correct train time table information right after the first query.

For all these reasons, much care had to be put on the design of the system. Additionally, to be aware of any other unforeseen problem, each recording session can be accompanied by a supervising person that knows about the structure of the simulation system and can guide the user in the right direction.

The first and very important step to guide the user into the predefined dialog is to start each dialog with a train time table inquiry given on the screen to be read by the user, e.g:

U: *"I want to go to Hamburg."*

From these first queries the tokens for the default accents (→ target_D1) were collected. After this query different system reactions are possible (cf. Figure 1), each of them provoking a specific user reaction. In the first situation (1a) a correct recognition of the user's query is simulated and the requested information, i.e. the correct train connection, is given. This provokes no specific user reaction but grants him/her a feeling of success. In (1b) a correct recognition is simulated, asking the user for confirmation. The usually following single word utterances (e.g. *"yes"*) or any other type of confirmation, can be collected as a by-product and re-used for training.

The best way to provoke the user to put stress on a specific syllable is to simulate recognition errors. In (2a) the user is induced to produce a contrastive accent on the lexical word accent position (usually he/she's going to utter: *"To* **Ham***burg"*). These utterances are used to collect the first type of contrastive accent (→ target_C1). With system reaction (2b) a contrastive accent on a specific syllable different from the word accent syllable can be provoked (induced user utterance: *"No, to Ham***burg"**). In this case the stress is put on the second syllable of the word (→ target_C2). With system reaction (2c) the user is induced to use a very distinct (emphatic) pronunciation where sometimes both syllables (→ target_C12) are overemphasized (esp. if this situation is used several times subsequently). Note that this mode of provoking accents was not used for the words examined in the following.

EXPERIMENTS AND RESULTS

Using the simulation system 15 recording sessions with 15 different users (180 dialogs in total) were conducted for collecting different types of accentuations, where all the intended minimal pairs comprised city names (like *"Hamburg"*, *"Freiburg"*) or time expressions (like *"at nine o'clock"*). Most of the users were students from the computer science department with no special knowledge of speech recognition or the EVAR system. They were told that their task is to test the automatic speech understanding system, and for the sake of convenience for the transcriber the first user utterance has to be read from the screen. At the end of each recording session, the users were asked about their experience with the system. None of them had any doubt that he/she was working with an automatic dialog system; most of them were very surprised about the systems capabilities and the computational speed.

In total 205 word tokens were collected, recorded and digitized using a *Desklab 14* from *Gradient*. Most of the tokens (62) were obtained for the city name *Hamburg*; in the following discussion, we confine ourselves to these items. The tokens were cut out of the signal, the syllable boundaries were adjusted by automatic time-alignment using an HMM-based word recognizer and corrected manually.

In an informal perceptual evaluation it was checked that the induced accentuation types were produced in the expected manner. Only 6 % of the induced contrastive accents were perceived as default accent; none of the default accents was perceived as a contrastive accent.

For the investigation of the prosodic properties of the different induced accentuation types, F0-contour and rms-energy (frame length: 10 ms) were computed automatically using the algorithm described in [3]. The F0-values were transformed into semi-tones. For F0 and energy the mean over the whole word was subtracted from each value. The following prosodic features were computed for each syllable: minimum, maximum, range, mean, onset and offset of the F0-contour; duration of the syllable nucleus; mean of the energy-contour.

Table 1: *Confusion matrix of induced and automatically classified accentuation types in percent.*

	# Tk.	D1	C1	C2
target_D1	28	78.6	10.7	10.7
target_C1	19	5.3	84.2	10.5
target_C2	15	13.3	26.7	60.0

In Table 1, the result of an automatic classification is shown (linear discriminant analysis, learn = test, all features used in a forced entry design). At first sight, the low recognition rate for target_C2 might surprise: 60% correct, and 26.7% confusion not with the default case target_D1 but with target_C1 where an 'opposite' accent pattern is expected. Of course, misproductions cannot be ruled out altogether and might – esp. if the number of tokens is as low as in our case – heavily influence the classification results. A systematic explanation along the lines of [2] can, however, be offered. There, a double focus on two different words was induced by the context but often it was classified and perceived *not* with focal accents on these two words but with one single accent on the word in the default ("out of the blue") accent position. But that means that speakers who do not "behave properly" – i.e. as the linguist likes them to do – do nevertheless deviate in a systematic manner. The same might be the case with contrastive accents: The strategy of naive speakers when confronted with a "contrastive misunderstanding" (*Ham***berg"**) instead of *Hamburg*) might sometimes be simply to repeat the word in question more pronounced in an overall manner but *not* – or not only – with a contrastive accent on the misunderstood syllable. As far as this behavior is representative for real life applications, it must be accounted for in the system.

In Table 2 the average of the feature values for both syllables is shown for the three induced classes. The duration of the syllable nucleus is most significant for

Table 2: *Average feature values for the three induced classes.*

Feature	target_D1		target_C1		target_C2	
# Token	28		19		15	
	Syl. 1	Syl. 2	Syl. 1	Syl. 2	Syl. 1	Syl. 2
Nucleus duration	153.6	119.2	192.4	156.0	201.0	176.5
F0–Mean	-0.15	0.37	0.40	-0.43	0.55	-0.55
F0–Maximum	1.61	1.75	2.16	1.32	2.60	1.60
F0–Minimum	-2.04	-1.25	-1.42	-2.05	-1.60	-2.93
F0–Range	3.64	3.00	3.58	3.37	4.20	4.53
F0–Onset	-0.39	0.11	-0.63	0.79	-0.53	0.13
F0–Offset	0.00	-0.07	0.89	-1.53	0.33	-1.87
Energy–Mean	-7.34	12.49	6.16	-4.48	-2.41	2.74

distinguishing default from contrastive accent; the tokens with contrastive accent are clearly longer than the default accents. The ratios between first and second syllable for default accent (1.29), contrastive accent on the first syllable (1.23) and contrast on the second syllable (1.14) moves towards a comparatively longer second syllable with the weakest differences in total syllable nucleus duration for target_C2. Still, the mean value of the absolute duration of the first syllable is for target_C2 slightly longer than for target_C1 and this fact corroborates our hypothesis that contrastive accentuation is not strictly refined to the syllable in question. The difference between the F0 features is not that distinct. The F0-range on the second syllable is clearly smaller for the default accent; the F0-mean, however, rises from the first to the second syllable. The energy proportions between first and second syllable show high differences for all three accentuation types. For the contrastive accents, these differences are as expected: higher energy on the accentuated syllable. For the default case, it is the other way round. Possible reasons might be that target_D1 was embedded in a complete sentence whereas the contrastive accents were usually just one word utterances and that no phoneme intrinsic normalization was performed for the energy.

The same features were extracted also for the automatically determined (not manually corrected) syllable positions. Same tendencies in the feature behavior could be observed, the differences were, however, less distinct.

CONCLUDING REMARKS

It has been shown that with the system described here, an automatic collection of contrastive accents produced in a natural way can easily be performed. Not only contrastive accents can by provoked with the system but, with some slight modifications of the system design, also other spontaneous speech phenomena like hesitations. Furthermore, preliminary experiments have already been conducted for the collection of spontaneous speech phenomena with the so called *"shocking effect"*, where an absolutely unexpected system answer like *"Why do you want to go there?"* is provoking very surprised user reactions.

ACKNOWLEDGEMENT

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grants 01 IV 102 F/4 and 01 IV 102 H/0. The responsibility for the contents lies with the authors.

REFERENCES

[1] R. Bannert. Fokus, Kontrast und Phrasenintonation im Deutschen. *Zeitschrift für Dialektologie und Linguistik*, Vol. 52, pp. 289–305, 1985.

[2] A. Batliner, W. Oppenrieder, E. Nöth, and G. Stallwitz. The Intonational Marking of Focal Structure: Wishful Thinking or Hard Fact? In *Proc. XIIth ICPhS*, Vol. 3, pp. 278–281, 1991.

[3] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP-Based Determination of *F0* contours from speech signals. In *Proc. ICASSP'92*, Vol. 2, pp. 17–20, 1992.

[4] W. Wahlster. Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. EUROSPEECH'93*, "Opening and Plenary Sessions", pp. 29–38, 1993.