BMBF

Verbmobil
Verbundvorhaben

FILLED PAUSES IN SPONTANEOUS SPEECH

A. Batliner, A. Kießling,
S. Burger, E. Nöth

L.M.-Universität München

F.–A.–Universität Erlangen–Nürnberg

Vm

Report 88
Juli 1995

Juli 1995

A. Batliner, A. Kießling,
S. Burger, E. Nöth

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

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

Tel.: (089) 2180 - 2916
e-mail: `ue102ac@cd1.lrz-muenchen.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 F/4 und 01 IV 102 H/0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

To be published in Proc. ICPhS'95:

FILLED PAUSES IN SPONTANEOUS SPEECH

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

ABSTRACT

Filled pauses as, e.g., uh, eh, signal disfluencies, i.e. hesitations or repairs. They do normally not occur in read speech and were therefore up to now rather seldom investigated; they must, however, be accounted for in the (automatic) processing of spontaneous speech. We present descriptive statistics and the results of an automatic classification of filled pauses in the database of the VERBMOBIL project and discuss the relevancy of different prosodic features for the marking of different types.

INTRODUCTION

Filled pauses (henceforth FPs) as, e.g., *uh, eh*, signal disfluencies and can be classified into

(1) **Hesitations** (FPHs) that are due to planning, control of turn taking, or speaker idiosyncrasies. Functional equivalents are unfilled pauses and hesitation lengthening that is not caused by accentuation or normal preboundary lengthening.

(2) **Cue phrases** (edit signals) for repetitions or repairs of words and phrases, or for restarts of syntactic constructions (FPRs). Functional equivalents are words like *no, that means*, etc. Often, such disfluencies are not marked by cue phrases but only with prosodic means.

Basically, the processing of FPs in human perception/comprehension and in automatic speech processing is analogous: FPHs should be disregarded with respect to linguistic content, FPRs can be taken as cues for a new parse where not only the FP but the reparandum as well has to be disregarded. A full account of these phenomena is given in [1]. In word recognition, FPs are usually only modelled as a waste paper basket category and disregarded. They are often confused with other words. More important than an improvement of word recognition might, however, be the use of FPs for higher linguistic modules as indication of different kinds of phrase boundaries, as an indication for the necessity to start a new parse, etc. It is not likely that FPs can be classified reliably only with spectral features. Several prosodic features are, however, reported in the literature as being relevant for the marking of FPs in English, cf. e.g. the results of [3] and [4]: The F0 of FPs is lower than that of the context, the restart after a FPR is often more stressed than the reparandum before the FPR, FPs at major boundaries are longer than within syntactic constituents.

MATERIAL AND PROCEDURE

Our material was recorded at four different sites for the spontaneous German database of the VERBMOBIL project (domain of appointment scheduling [5]). Because of inconsistencies in the rest of the material, only data recorded at the two sites Karlsruhe and Munich will be used. In total, 2422 turns (339 minutes of speech) from 56 female and 81 male speakers were investigated.

In the basic transliteration, there are four different types of FPs with the following tokens given in SAMPA notation:

<äh>	6:, 6, E:, E, @:, @, 2, 9
<ähm>	6:m, 6m:, 6m, E:m, Em:, Em, @:m, @m:, @m, 2m:, 2m, 9m:, 9m
<hm>	hm, hm:, m, m:
<häs>	pu, pu:, f, f:, pf, pf:, ...

In the transliteration, FPRs can easily be distinguished automatically from FPHs because the disfluencies in their vicinity are labelled separately. The distribution of the four types of FPH and FPR within the 2422 turns is given in Table 1 together with their sum (FP) and, so to speak, their functional complement ($\mathcal{C}$). There, either a <Z> denotes a lengthening of the final syllable in a word that is not only caused

Table 1: *Distribution of* FP*s*

	äh	ähm	hm	häs	FP	$\mathcal{C}$
hesitations	471	368	59	70	968	964
repairs etc.	63	23	7	8	101	483

by a following higher syntactic boundary (i.e. 'regular' preboundary lengthening) or repetitions/repairs/restarts are found without FPs. 4% (35 cases) of the FPs are adjacent to <Z>, and 3% (25 cases) to pauses (<P>, 687 tokens) that are labelled if a clear silent interval of more than 0.5 sec can be perceived; 35% (337 cases) of the FPs are adjacent to breathing (<A>, 3001 tokens). 'Adjacent' in this context means 'strictly adjacent' i.e. not separated by any other event. Hesitations are thus almost always signalled either by FPH or by <Z> but not by both. Breathing cooccurs very often with higher syntactic boundaries and thus also with FPs at these boundaries. In the average, almost every second turn or every 19th sec, a FP can be observed. FPs amount to 2% of the vocabulary; in comparison, the most frequent word *ich* amounts to 3%; ca. 85% of the FPs are <äh> and <ähm>. FPHs are roughly ten times more frequent than FPRs. No gender specific difference could be observed as for average length of turns or overall frequency of FPHs or FPRs.

For the prosodic characterization, we used a large set of 47 syllable based features similar to those that proved to be relevant for the automatic classification of phrase boundaries and accents [2]: Duration: (dur) in ms and normalized (durno) as in [2]; for energy ("loudness"): mean (enmean), median (enmed), maximum (enmax), regression coefficient (enreg), and squared mean error of the regression coefficient (enerr); for F0, normalized with respect to range (logarithmized) and utterance (mean of utterance subtracted): mean (F0mean), median (F0med), maximum (F0max), regression coefficient (F0reg), squared mean error of the regression coefficient (F0err), minimum (F0min), onset (F0ons), and offset (F0off); length of pause (pause) before and after the (FPs). The features

Table 2: *Percentage of* FP*s at boundaries*

type	position	%
Wi	word internal	0
B0	any other word boundary	6
B1	constituent boundary	25
B2	weak/intermediate boundary	15
B3	strong/phrase boundary	31
Ti	turn initial	14
Tf	turn final	0
R	repair/restart/repetition	9

were extracted for three syllables before the FP (Index $\overline{31}$), the FP itself (Index $\overline{00}$), and three syllables after the FP (Index $\overline{13}$). These features are of course often highly correlated with each other. Their combined use, however, prevents from excluding features that are more relevant than those that might have been chosen by purely phonetic reasoning.

The position of syllable boundaries was computed by an automatic time alignment using a HMM based word recognizer. F0 and energy features were extracted automatically. For paradigmatic comparison, two control syllables with similar phonetic shape were processed as well: [vEm] in *November*, 125 tokens, and [vE:r] in *wär'*, 185 tokens.

RESULTS AND DISCUSSION

In the following, we will disregard the waste paper basket category <häs> because of its varying phonetic substance, and combine the remaining three types. Table 2 shows the distribution of FPs for different positions; the very few Wi and Tf types will be disregarded as well: B0, B1, and B2 constitute the class **FPHweak** at weak, B3 and Ti the class **FPHstrong** at strong boundaries [2]. These types were labelled manually in the transliteration, B2 e.g. in the vicinity of a comma, B3 in the vicinity of a period or a question mark. Final correction of the punctuation in the transliteration and of the labelling of FP types was done by one of the authors; even if these labels are not strictly based on a linguistic analysis, they are thus fairly reliable.

A thorough discussion of the results is beyond the scope (and especially space) of this paper. We will only present the most evident and important findings that are

Table 3: *Automatic classifications*

	constellation of classes	feat.	%
(1)	(B0)(B1)(B2)(B3)(FPR)	$\overline{31},\overline{00},\overline{13}$	46
(2)	(B0 B1 B2)(B3 Ti)(FPR)	$\overline{00},\overline{13}$	55
(3)	(B0 B1 B2)(B3)(FPR)	$\overline{31},\overline{00},\overline{13}$	59
(4)	(B0 B1 B2 B3)(FPR)	$\overline{31},\overline{00},\overline{13}$	68
(5)	(B0 B1 B2)(B3)	$\overline{31},\overline{00},\overline{13}$	77
(6)	(B3)(Ti)	$\overline{00},\overline{13}$	84
(7)	(B0 B1 B2 B3 FPR) (vEm)(vE:r)	$\overline{31},\overline{00},\overline{13}$	85
(8)	(B0 B1 B2 B3 FPR) (vEm vE:r)	$\overline{31},\overline{00},\overline{13}$	91

based on an automatic classification (linear discriminant analysis) where all features were used in a learn=test, forced entry design. Overadaptation takes place with learn=test, and the percent correctly classified can therefore not be taken as a realistic estimate for real life application. We can, however, estimate the relevancy of the features looking at their correlation with the discriminant function, and we can estimate the difference in predictability between those constellations that are given in Table 3 that shows classes to predict, features used (feat.), and percent correct (%). Chance level for the five classes in (1) is 20%, for three classes 33% and for two classes 50%. For Ti-FPs, preceding context, i.e. $\overline{31}$–features, are not available. It was therefore necessary to either exclude these features as in rows (2) and (6) or to exclude this class as in the other constellations from the analysis.

All results in Table 3 are well above chance level. Promising are the results of (7) and (8) because they show that prosodic features really can help in telling apart FPs from other syllables, the most important feature being durno, cf. below. In the other analyses, fewer classes result in better classification; that could be expected because the chance level increases as well. We can doubt whether in real life applications, different types of FPs can be told apart with a reasonably high probability but in the long run, not only the prosodic features used can be fed into the analysis but other features as well; e.g. the presence of breathing, cf. above, makes it more likely that a FP belongs to FPstrong etc. Even a rather simple language model might be very useful as well. Another factor might be that the database so far is relatively small; more data will hopefully result in a better statistical modelling and thus in better classification rates. (Note that the influence of random errors that always are contained in automatically extracted feature values diminishes if more data are used.)

We want to discuss row (3) in more detail, where (B0 B1 B2), i.e. FPHweak, (B3), i.e. FPHstrong, and FPR are contrasted. FPR tends to be confused more with FPHweak than with FPHstrong and vice versa, pause being more pronounced for FPHstrong than for the two other classes. In Table 4, mean values are presented for the most relevant four cover classes and for most of the features apart from enmed, F0med, F0max, and F0min where the relevant information is mostly encoded in other features (mean values or range). For convenience, energy values apart from enreg are divided by 10, and F0range is multiplied by 100. If we look at these mean values and at the correlation of the features with the canonical discriminant function, we can, with due care, assume that pause, energy and duration features (in this order) are most important for contrasting FPHstrong from the other two FPs on the one hand, and on the other hand, that energy and F0 features, esp. F0reg$\overline{13}$ and F0reg$\overline{31}$, are most important for contrasting FPRs from the two FPH classes. That means that prototypically, FPHstrong has longer adjacent pauses than FPHweak or FPRs and less energy on the preceding syllables; this finding is plausible as higher syntactic boundaries are expected to be marked with pauses and with a final energy decline. For FPRs, the F0 regression line on the preceding syllables is more falling, and the F0 regression line on the following syllables is more rising than in FPHs. The energy on the adjacent syllables is lower in FPRs than in FPHs. It might not surprise that energy on the following syllables is lower for FPRs than for FPHs even if usually, it is assumed that the reparandum is more stressed than other syllables: energy

Table 4: *Mean values of relevant features for four cover classes*

type	FPHweak			FPHstrong (B3)			FPR			vEm ∪ vE:r		
context	$\overline{31}$	$\overline{00}$	$\overline{13}$	$\overline{31}$	$\overline{00}$	$\overline{13}$	$\overline{31}$	$\overline{00}$	$\overline{13}$	$\overline{31}$	$\overline{00}$	$\overline{13}$
pause	225	—	301	429	—	791	292	—	242	22	—	35
dur	108	215	87	104	193	79	111	201	81	77	65	77
durno	.83	2.32	.11	.59	2.46	.01	.77	2.09	-.00	.01	-.12	.11
enmean	324	367	329	298	359	344	296	354	308	296	414	309
enmax	720	500	705	620	522	744	671	487	679	643	641	651
enreg	-4.49	37.64	16.34	-5.05	-3.00	22.79	-5.82	81.02	22.69	7.81	158.97	-14.54
enerr	1685	416	1416	1370	480	1512	1566	348	1345	1285	536	1301
F0mean	.027	-.042	.017	.005	-.040	.030	.013	-.060	.026	.023	-.008	.010
F0ons	.022	-.026	-.020	-.000	-.032	-.026	.034	-.046	-.041	.031	-.017	.017
F0off	.010	-.045	.030	.013	-.043	.063	-.032	-.060	.058	-.012	.012	.013
F0range	.272	.085	.268	.294	.073	.241	.306	.083	.289	.243	.110	.238
F0reg	-.000	-.078	.070	.036	-.052	.179	-.097	-.117	.226	-.083	.132	-.010
F0err	.574	.104	.462	.574	.090	.409	.636	.090	.519	.413	.095	.367

might be less important for accentuation than duration or F0 features, e.g. the rising F0 regression line after FPRs.

If we compare FPs with the control syllables, the most important feature is duration, regardless whether it is normalized or not. This might be due to the fact that the control syllables are intrinsically rather short, and that we simply have chosen "biased" control syllables. But even without all durational features, classification is only ca. 5% worse than with durational features. That means that the other features encode enough relevant information, most important being the adjacent pauses that are way shorter for the control syllables than for the FPs. F0 values are lower and F0 regression line is more falling in FPs; this finding corroborates the hypothesis that FPs behave like parenthetical chunks that have lower F0 than their surrounding.

CONCLUDING REMARKS

The results achieved for our spontaneous German database are similar to those of e.g. [3] and [4] where English material was investigated. (They are, however, not identical: in contrast to [3], FPHstrong is, e.g., not longer than FPHweak.) We didn't have a close "phonetically minded" look at some selected features but have tried to include a very large set of prosodic features. The picture that emerges from this data driven approach is possibly more complicated than expected; it is e.g. rather difficult to judge and to explain the relevancy of the different energy features. More data is needed and more space to disentangle matters. But we can expect that very large databases are available in the near future and we hope that with such an approach, the epistemological gap between knowledge based methods (phonetics) and statistically based methods (automatic speech processing) will diminish in the long run.

ACKNOWLEDGEMENTS

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] A. Batliner, S. Burger, and A. Kießling. Außergrammatische Phänomene in der Spontansprache: Gegenstandsbereich, Beschreibung, Merkmalinventar, Verbmobil–Report, Nr. 57, 1994.

[2] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. Detection of Phrase Boundaries and Accents. In Niemann, de Mori, and Hanrieder, editors, *Progress and Prospects of Speech Research and Technology*, infix, Sankt Augustin, pp. 266–269, 1994.

[3] D. O'Shaughnessy. Locating Disfluencies in Spontaneous Speech: An Acoustic Analysis. In *Proc. EUROSPEECH'93*, Vol. 3, pp. 2187–2190, Berlin, 1993.

[4] E.E. Shriberg and R.J. Lickley. Intonation of Clause-Internal Filled Pauses. *Phonetica*, Vol. 50, pp. 172–179, 1993.

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