Cognitive aspects of communicating information with conversational fillers in Slovak

Štefan Beňuš
Constantine the Philosopher University, Nitra, Slovakia
Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
sbenus@ukf.sk

***Abstract*—This paper reports on the *form-function* relationship of Slovak conversational fillers *uh* and *mm* in task-oriented dyadic spontaneous human-human conversations. The form is represented by two phonetic features: a discrete feature of nasality and a continuous feature of duration. The function is assessed on three dimensions critical for human-machine communicative systems: turn-taking management, (meta)cognitive state, and discourse and informational structure. We report that the two phonetic features facilitate partial disambiguation of communicative functions and discuss potential for applicability of our observations for human-machine interactive voice systems.**

Keywords—conversational fillers; turn-taking; human-machine interaction

I. Introduction

Conversational fillers (CFs) are sounds commonly transcribed as *uh*, *um*, *err*, *mm*, *ah,* or other. They are typically realized with mid-central schwa-like vowel and a nasalized bilabial component but both of these are optional. They are ubiquitous in spontaneous conversational speech reaching most commonly rates around 2-5% of all words ([1], [2], [3]). One of their primary pragmatic functions is to signal cognitive processing. In dialogues, this cognitive function might be bi-directional. A CF might be "forward-looking" and signal planning difficulties associated with a choice in producing the current utterance by directing attention to the material following the CF; see [4] for a recent review. CFs also offer a peek for the listeners into metacognitive processing of the speakers. If a turn follows a question and starts with a CF, listeners tend to expect that the speaker knows the answer [5].

Especially in turn-initial positions CFs might also be "backward-looking" and point to difficulties the current speaker has in parsing and/or comprehending the utterance in the preceding turn. In this sense, for example, the speaker signals uncertainty [6] and thus need for further elaboration. However, this forward-backward looking dichotomy is often difficult to maintain since many turn-initial CFs might simultaneously communicate information in both directions. For example, [7] proposed that some uses of turn-initial CFs signal both acknowledgment of information in the previous turn, that commonly requests information, and planning how to provide and package the requested information.

In addition to (meta)cognitive functions, CFs have also been found to communicate several discourse functions related to floor management and, by extension, to turn-taking, as well as to discourse/information structure of dialogues. In her analysis of the Lund corpus of English, [8] found that fillers mark speakers' intentions to assume and hold the floor in dialogues. CFs are also necessary in managing spontaneous-like conversations [9], [10] and thus, in general, should be "understood as devices with important turn-organizational uses" [11], p.720. Regarding discourse structure, reference [12], in his analysis of spoken Dutch, argued that phrases following major discourse boundaries tend to contain CFs. Fillers in phrases that followed major discourse breaks tended to occur phrase-initially whereas CFs that followed minor breaks were found mostly in the phrase-internal position. CFs at major discourse boundaries were both segmentally and prosodically distinct from those found at minor boundaries. Speakers tended to signal major breaks with *ums* rather than *uhs*, and they produced them with higher fundamental frequency and longer duration. Hence, a long *um* increases chances that a new discourse segment/topic will follow, which is very useful information for automatic dialogue systems.

Another discourse dimension is the distinction between given vs. new discourse referents. For example, [13] showed that listeners expect discourse-new referents when hearing disfluent instructions such as those including fillers. Additionally, [14] observed that *um* was four times more likely to precede new referents than *uh* in production but not in perception. Although the type of filler did not affect the perception of given vs. new, the presence of a filler in itself did signal new referents. This might be related to transitional probabilities since [15] argued that CFs warn listeners that a low-probability transition word is following.

Regarding the differences between CF types, the most common fillers *uh* and *um* in (American) English received significant attention. References [16] and [17] suggested that despite their lexical "emptiness", CFs should be considered words since *uh* signals a short interruption or delay while *um* signals a more serious longer one. Reference [18] similarly argued that *um* is more common initially and reflects thus "planning of larger units, while 'uh' may be relatively more likely to reflect local lexical-decision making" (p. 154). Other studies reported no discernible difference between the two CF types, e.g. in expectation that a speaker knows the answer to a question [5].

Given this extensive research into the cognitive and interactional aspects of CF production and perception in

978-1-4799-1546-0/13/$31.00 ©2013 IEEE

English, it is warranted to investigate situation in other languages. Some studies, of mostly Germanic languages exist [12], [19], but wider array of languages provide more robust findings. The overall goal is to discover which aspects of form-function relationship involving CFs are cross-linguistically valid and which display language-, or culture-, specific features. Improved understanding of these aspects brings great potential for improving the quality and naturalness of human-machine interactions through dialogue systems and interactive voice response applications. Moreover, although some applications using CF in speech generation exist, e.g. [20] showed that a turn-initial CF produced by a robot significantly improved the user's impression of longer response times, the design of cognitive artificial agents, equipped with limited functionality in recognizing and producing the form-function relationship associated with CFs, still represents a challenging goal. In this paper we attempt to pursue these overall goals by reporting on the form-function relationship of Slovak conversational fillers *uh* and *mm* in task-oriented dyadic spontaneous human-human conversations.

II. Methodology

A. Corpus

Data for this study comes from a corpus of dyadic conversational games inspired by Columbia Games Corpus [21], [22], [23]. The Slovak version of this corpus is described also in [24]. Briefly, two subjects were seated facing a monitor but without visual contact between each other. In the game, they earn points for positioning various objects on their computer screens with a mouse. Crucially, the game required spontaneous spoken interaction in which one player described the position of the target object with respect to other objects on the screen and the other player tried to place the target object as close as possible to this target position. In each session, the subjects placed 14 objects alternating their roles of Describer and Placer.

The corpus was manually transcribed and the text was then automatically forced aligned to the acoustic signal using HVite utility from Hidden Markov Model Toolkit [25]. Subsequently, this alignment was manually checked and corrected by the author.

In this study we analyze a subset of the corpus consisting of 5 sessions played by 7 speakers (3 females and 4 males); 3 speakers played the game twice with a different partner. In this subset, there are 7.1 hours (425.2m) of speech, 24919 words, out which 658 (2.64%) are conversational fillers.

B. Labeling

A novel pilot labeling scheme was designed for investigating the relationship between the forms and functions of conversational fillers in the corpus of interactive interpersonal task-oriented dialogues. Following the review of functions in Section I, we took three core functional dimensions of CF usage: interactional, cognitive, and informational. The interactional function relates to the management of turn-taking between interlocutors. Here, fillers most commonly signal the wish to hold the floor and continue speaking, (H)OLD in TABLE I. Very commonly, conversational fillers also appear turn-initially. In this position they may either express the desire of the speaker to grab the floor (G)RAB, if it was not pre-selected to him/her [11], or, simply assume the floor if the speaker was selected (I)NITIATE. Finally, a filler might signal a wish to yield the floor to the interlocutor (Y)IELD, serving as a prompt for more input from the interlocutor, e.g. [3].

The second dimension reflects (meta)cognitive processing of the speaker. In this we employ the hypothesized difference in planning difficulty between problems with lexical access of a target entity (e.g. picture frame), and deeper problems with planning larger cognitive chunks of information. The former appears as (L)EXICAL ACCESS in TABLE I. and the latter as (P)LANNING. Fillers are also commonly used to signal an upcoming editing of previously given information when speech errors, repeats, re-start, or self-corrections appear. In this area, we differentiate between an error or repetition linked to a single lexical item, denoted by (E)RROR in TABLE I. , and a re-start or re-formulation of an idea relating to a larger informational chunk; (C)ORRECTION.

The third functional domain attempts to tap into the discourse structure as signaled by conversational fillers. Here we employ two broad categories. The first is “backward-looking” and captures the uses in which a filler (R)ESPONDS to a preceding utterance from the interlocutor. This response may include an acknowledgment, readiness to answer a question, or evaluation of previous utterance or game activity. The second is “forward-looking” and signals that the upcoming information is worth paying attention to. It includes two sub-categories: introducing a discourse-new item; (N)EW, and beginning a new larger discourse segment; (B)EGIN SEGMENT.

TABLE I. LABELING SCHEME FOR FUNCTIONS OF CONVERSATIONAL FILLERS

Label	Dimension	Function description
H	Turn-taking	Hold the floor
G		Grab the floor
I		Initiate a turn when pre-selected
Y		Yield the floor
L	(Meta-) Cognitive	Lexical access problem
P		Planning a larger information chunk
E		Error, signal that previous item will be edited
C		Correction of larger information chunk, re-phrase
R	Discourse	Respond to previous utterance/activity
N		Upcoming Discourse-new item
B		Begin a new discourse segment, a new idea

If none of the functions for a particular domain was possible to identify for a particular CF token, it was deemed not applicable to that dimension, and “X” for that dimension was entered. A single annotator (the author) labeled all CF tokens in the subset of the corpus described in Section II.A. All of these categories are rather broad and not unambiguous. The scheme is taken as a first effort to be adjusted if needed in subsequent labeling of the entire corpus.

III. Results

A. Descriptive observations

The overall mean rate of filler usage was 2.64%, which corresponds well with the rates of CFs in other spoken spontaneous corpora (e.g. [1], [2]). The speakers varied between the minimal CF use by RS (1.1%) and maximal use by KM (5.2%). CFs in Slovak thus, similar to spoken corpora in other languages, belong among the most frequent words.

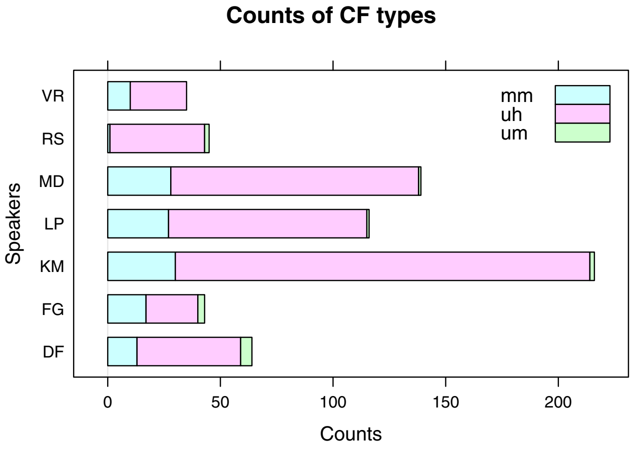

Fig. 1. Counts of three conversational fillers divided by speakers.

Regarding the types of conversational fillers, Fig. 1 illustrates the data. We see that for each speaker and overall, the most frequent CF is *uh*, significantly less frequent is *mm* and *um* is extremely rare with only 14 tokens in total. In the following, we pool the two nasalized variants (*mm* and *um*) together as *mm*.

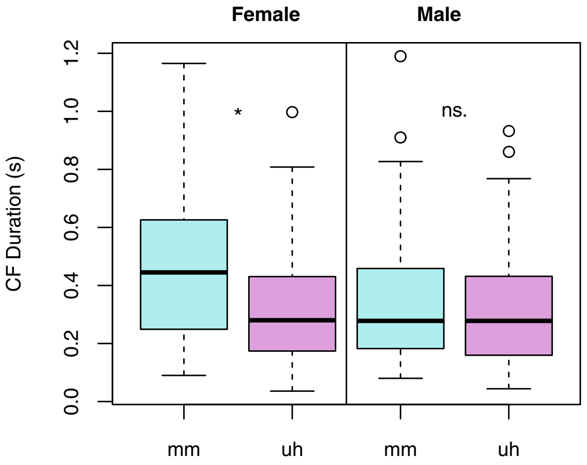

Fig. 2. Duration of conversational fillers split for gender and CF type.

One of the observations from English is that a non-nasalized CF is shorter than the nasalized one and also that silent pauses adjacent to the former are shorter than to the latter. These were taken to support the view that *um* signals deeper cognitive planning problems than *uh*. In our data, as shown in Fig. 2, this applies only to female speakers. A mixed-models test with speaker as a random factor showed a significant effect of CF type and a significant interaction with Gender. In separate tests, *mm* was significantly longer than *uh* only for females (F = 29.3, p < 0.001) and not for males (F = 0.9, n.s.).

Another observation concerned the duration of silent intervals adjacent to the fillers. We do have labeling of long major silent pauses, these might be both turn-internal and across turns, and short minor ones that are by definition always turn-internal. For *mm*, indeed, 50% of the following intervals are long silent pauses and only 5% are short ones. For *uh*, only 32% are long pauses and 10% are short ones. Hence, compared to *uh*, *mm* is more likely to precede a silent pause than a word, and these silent pauses are more likely to be long than short ones.

We tested this observation from the discrete labeling also by examining the relationship between continuous intervals of CF duration and the duration of a following silent pause. Fig. 3 shows this relationship separately for the two CF types. The first observation is that the relationship is not very robust as the slopes of the regression lines are very flat. Second, the linear regression tests show that the positive correlation between the two intervals is significant for *uh* (t = 4.76, p < 0.001, R^2 = 0.09) and not significant for mm (t = 0.75, p = 0.46, R^2 = -0.01). This result was consistent irrespective of gender. The variability of Slovak *mm* findings suggests that, cognitively, it might be linked less to planning difficulties than English *um*. We will return to this suggestion below.

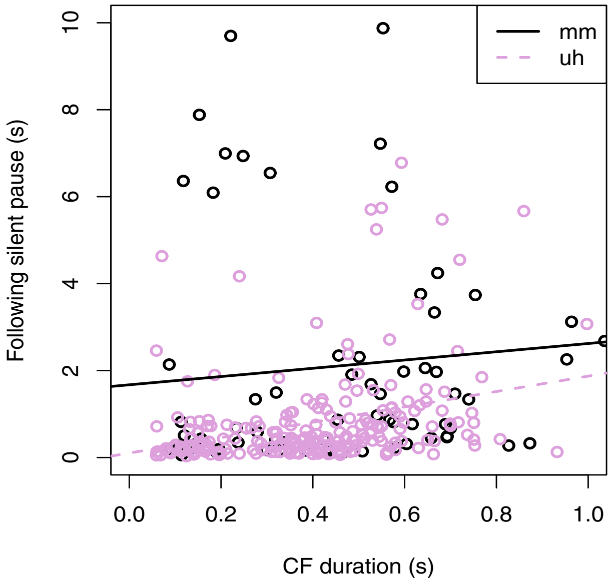

Fig. 3. Duration of the following silent pause (limited to 10 s.) as a function of CF duration..

B. *Labelled CF functions*

Following the labeling scheme introduced in Section II.B, we first describe the three functional dimensions in turn and then discuss the overall patterns. Considering the first dimension of turn-taking management, TABLE II. summarizes the data. A chi-square test reports a significant difference between the two CF types (X^2 [4] = 39.0, $p < 0.001$) but the subsequent analysis of the residuals shows that the significant difference obtains only for the "H" and "I" labels: *uh* is more frequent for the floor holding function than *mm*, and the opposite applies to initiating a turn after being pre-selected. Shading indicates these significant differences (pink for the former and light cyan for the latter). Hence, *mm* seems more suitable for initiating turns while *uh* for holding them.

TABLE II. RATES OF TURN-TAKING FUNCTIONS SEPARATELY FOR MM AND UH; SHADED ROWS INDICATE SIGNIFICANT RESIDUALS AT P < 0.05.

Function	mm		uh		Total	
	N	*%*	*N*	*%*	*N*	*%*
Hold	72	51.4	395	76.3	467	71
Grab	8	5.7	27	5.2	35	5.3
Initiate	56	40	87	16.8	143	21.7
Yield	0	0	2	0.4	2	0.3
X	4	2.9	7	1.4	11	1.7

Nevertheless, the table shows that both CF types can also be used for the alternative function. Given this overlap, and the duration difference between the two CFs (for females) reported in Fig. 2, we hypothesized that CF duration might be another cue for differentiating the "Hold" and "Initiate" functions for turn-management. Separate Welch two sample t-tests for the two CF types support this hypothesis only for *uh* (t(127.3) = -2.48, p = 0.014) but not for *mm* (t(119.2) = 0.74, p = 0.46). Hence both *mm* and longer *uh* tokens are associated with turn-initial position while short *uh* tokens signal holding the floor for the speaker.

We now turn to the second (meta)cognitive dimension of the labeling scheme with summarizing the crucial information. A chi-square test reports a significant difference between the two CF types (X^2 [4] = 79.9, $p < 0.001$) but the subsequent analysis of the residuals shows that the significant departure from the expected values obtains only for lexical access, error and also the "X" label indicating no discernible function. As the percent rates and shading of the table indicate, *uh* is more frequent than *mm* for signaling minor planning problems commonly associated with lexical access, and for emitting an edit signal indicating an upcoming local correction or repetition of previous material.

Regarding the question if CF duration might help in disambiguating these functions, Fig. 4 illustrates the data. An Anova test with CF type (*uh* vs. *mm*) and (meta)cognitive function (L vs. P vs. E vs. C vs. X) as factors and CF duration as a dependent variable reports a significant effect of (meta)cognitive function (F(1,4) = 14.3, $p < 0.001$) and no significant interaction between the two factors. Hence, we directly moved to post-hoc pair-wise testing using the Bonferroni adjustment. These tests showed that the longest CFs signal major planning issues ("P"), and these CFs are significantly longer than those signaling minor lexical access problems ("L"), local speech errors ("E"), and corrections/re-formulations ("C"). The significance is marked with a "*" in the figure. We see, however, that for the last two, the significance comes from *uh* tokens, that are also more numerous than *mm* ones as seen in TABLE III. Finally, no significant difference was reported for L-E and L-C pairs. To sum up, *uh* is typically used for signaling lexical access problems and errors, and the longest CFs signal major planning issues related to formulating and packaging information.

TABLE III. RATES OF (META-)COGNITIVE FUNCTIONS SEPARATELY FOR MM AND UH; SHADED ROWS INDICATE SIGNIFICANT RESIDUALS AT P < 0.05.

Function	Mm		uh		Total	
	N	*%*	*N*	*%*	*N*	*%*
Lex. access	10	7.1	130	25.1	140	21.3
Plan	92	65.7	291	56.2	383	58.2
Error	3	2.1	44	8.5	47	7.1
Correction	7	5	40	7.7	47	7.1
X	28	20	13	2.5	41	6.2

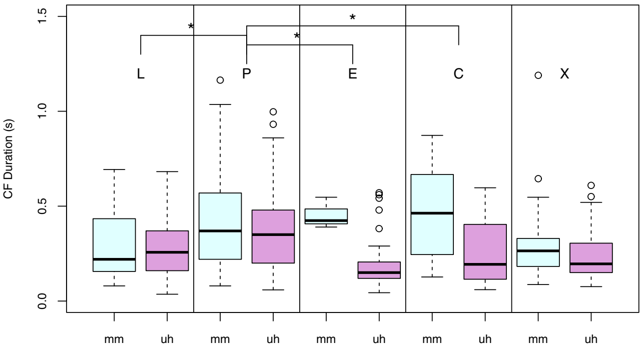

Fig. 4. Duration of conversational fillers as a function of metacognitive function and CF type; see text for further details.

The last third dimension of labeling concerned discourse and its structure, and the data are summarized in TABLE IV. that lists the counts and rates. The distribution again significantly departed from the expected one (X^2 [3] = 77.3, $p < 0.001$). This dimension is different from the previous two in that more than half of the tokens were not assigned a label ("X"), and this was significantly more common with *uh* than with *mm*. Additionally, *uh* is likely to signal an upcoming discourse-new item. CF *mm*, on the other hand, was significantly more frequent in responding to previous utterance or action showing thus the evaluative function of this CF, to which we will return in the discussion.

TABLE IV. RATES OF DISCOURSE FUNCTIONS SEPARATELY FOR MM AND UH; SHADED ROWS INDICATE SIGNIFICANT RESIDUALS AT P < 0.05.

Function	mm		uh		Total	
	N	*%*	*N*	*%*	*N*	*%*
Respond	55	39.3	52	10	107	16.3
New disc. item	2	1.4	61	11.8	63	9.6
Begin disc. seg.	28	20	105	20.3	133	20.2
X	55	39.3	300	57.9	355	54

Regarding testing for possible interaction between CF type and discourse labels on CF duration similar to previous two dimensions, no consistent result can be reported. Despite significant interaction between the two factors in an Anova test (F(1,3) = 5.5, p = 0.001), discourse labels did not affect CF duration significantly neither in the pooled data (F(1,3) = 1.1, p = 0.36) nor separately for two CF types where only tendencies were reported (F(1,3) = 2.5, p = 0.057 and F(1,3) = 2.5, p = 0.062 for *mm* and *uh* respectively).

Finally, to analyze the overall patterns, Table V shows the most frequent three-label combinations in the corpus. These appeared more than 10 times (1.5%) and together represent 91.5% of all tokens. The most frequent use of CFs is to signal minor or major planning issues in turn-internal positions (HPX, HLX), which together comprises almost 40% of all tokens. The second observation complements the analysis of differences between the two CF types for three dimensions separately. We see that *mm* commonly appears in the turn-initial position to respond to previous utterance or actions and/or express evaluative comment. In this function, *mm* dominates *uh*. Conversely, *uh* tends to signal turn-internal minor lexical access problems, errors, and discourse-new items.

TABLE V. RATES OF 3-LABEL COMBINATIONS SEPARATELY FOR MM AND UH; SHADED ROWS INDICATE A SIGNIFICANT CHI-SQUARE TEST AT P < 0.05.

Label	mm		uh		Total	
	N	*%*	*N*	*%*	*N*	*%*
HPX	38	27.1	128	24.7	166	25.2
HLX	7	5	86	16.6	93	14.1
HPB	9	6.4	45	8.7	54	8.2
IPR	23	16.4	29	5.6	52	7.9
HEX	3	2.1	42	8.1	45	6.8
HLN	2	1.4	40	7.7	42	6.4
IPB	9	6.4	32	6.2	41	6.2
HCX	5	3.6	26	5	31	4.7
IXR	21	15	3	0.6	24	3.6
GPB	4	2.9	12	2.3	16	2.4
GPR	3	2.1	12	2.3	15	2.3
HPN	0	0	12	2.3	12	1.8
HCB	2	1.4	8	1.5	10	1.5

IV. SUMMARY AND DISCUSSION

We set out to analyze the relationship between two phonetic features of Slovak conversational fillers, their duration and nasality, and their pragmatic functions relating to three dimensions: turn-taking management, meta-cognitive processing, and discourse and information structure. To summarize our main observations, we reported robustly greater frequency of non-nasalized CFs compared to the nasalized ones, and within the second category, wide-spread use of *mm* and extremely limited use of *um*. Regarding CF duration, we expected longer *mm* CFs than *uh* ones but this expectation was born out only for female speakers. We also hypothesized that CFs signal to the interlocutor cognitive states of planning. The analysis of material following CFs corroborates this CF function. We found that *mm* is more likely to precede a silent pause than a word, and these silent pauses are more likely to be long than short ones. However, positive correlation between the length of CF and the length of subsequent silent pause was significant only for *uh* tokens indicating that *uh* is cognitively a more robust signal of planning than *mm*. This observation was supported by analyzing the pragmatic functions of CFs and their relation to duration. We saw that *mm* CFs feature commonly as turn-initial evaluative or attitudinal comments, or introduce such comment, in which they respond to or acknowledge previous utterance or action. The non-nasalized *uh* tokens, on the contrary, typically express the wish of the speaker to hold the turn, to re-phrase or correct what was previously said, or introduce a focused item important for the current discourse.

Despite these observations, Slovak CFs remain relatively overloaded due to massive overlaps in the pragmatic functions signaled by the two CF types, which applies more to *uh* than to *mm*. Here, the phonetic duration of CFs might provide additional cues for increased success in disambiguating these functions. Specifically, the longer the CF, the more likely it is to signal turn-initiation and global cognitive planning of larger information chunks rather than local signals relating to previous items to be corrected or focusing the following items.

There are at least two avenues how this improved understanding of the relationship between phonetic form and pragmatic meanings of Slovak conversational fillers in task-oriented dialogs between humans may aid efforts to building human-machine applications ready for cognitively based information transfer that is communicatively natural and emotionally colored. The first avenue relates to cross-linguistic research of these communicative signals. Comparing observations from languages such as English and Slovak leads to uncovering those aspects of form-function relationship that are cross-linguistically valid, and thus applicable within a general interactional module. Such a module might provide basic human-machine functionality for any language. In such an approach, this basic language-independent interactional module would be connected to, but in principle independent from, a dedicated language-specific module. This relates to the efforts for building suitable computational paradigms and architectures for testing and implementing these ideas in human-machine interactive devices using speech as the primary modality.

The second avenue involves increasing robustness of future applications. The machines are already required, and will be more and more in future, to "sense" the attitudes, emotional states, and stances of people toward various communicatively meaningful acts. Since the expression of these is marred with ambiguity, partial information obtainable from frequent conversational fillers should improve interactive voice-based communication between human and machine. Moreover, some information can be obtained "fast and cheaply" from the signal, such as observations in Section III.A. Data for this are basically determined from the signal after employing accessible NLP technologies. Similarly, if the machine's models of reality can be aided with better understanding of human communicative signals such as CFs, these models should be improved. Finally, implementing conversational fillers into the machines' spoken productions in line with pragmatic features discussed in this paper might increase naturalness of such speech. This is an important step for increasing trust and credibility of human-machine systems aiming thus at improving well-being of its users on the one hand, and wider reach and impact of these systems on the other.

Acknowledgment

This publication is the result of the project implementation: Technology research for the management of business processes in heterogeneous distributed systems in real time with the support of multimodal communication, RPKOM, ITMS 26240220064 supported by the Research & Development Operational Programme funded by the ERDF.

References

[1] E. Shriberg, "To "Errrr" is human: Ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association* 31(1), pp. 153-169, 2001.

[2] Š. Beňuš, F. Enos, J. Hirschberg, E. Shriberg, "Pauses and deceptive speech", in *Proceedings of 3rd International Conference on Speech Prosody*, Dresden, 2006.

[3] Š. Beňuš, "Variability and stability in collaborative dialogues: turn-taking and filled pauses," *Proceedings of the 10th INTERSPEECH*, pp. 709-799, 2009.

[4] O. W. Stewart, M. Corley, "Hesitation disfluencies in spontaneous speech: the meaning of um," *Language and Linguistics Compass* 4, pp. 589–602, 2008.

[5] S. E. Brennan, M. Williams, "The feeling of another's knowing: prosody and conversational fillers as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language* 34, pp. 383–398, 1995.

[6] J. Fox Tree, "Interpreting Pauses and Ums at Turn Exchanges," *Discourse Processes* 34(1), pp. 37-55, 2002.

[7] Š. Beňuš, A. Gravano, J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, 43(12), pp. 3001-3027.

[8] A. Stenström, "Pauses in monologue and dialogue," In J. Svartvik (ed.) *London-Lund Corpus of Spoken English: Description and Research*, Lund: Lund University Press, 1990.

[9] H. Bortfeld, S. Leon, J. Bloom, M. Schober, S. Brennan, "Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender," *Language and Speech* 44(2), pp. 123-147, 2001.

[10] M. Taboada., "Spontaneous and non-spontaneous turn-taking," *Journal of Pragmatics* 16(2-3), pp. 329-360, 2006.

[11] H. Sacks, E. Schegloff, G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, 50, pp. 696–735, 1974.

[12] M. Swerts, "Conversational fillers as markers of discourse structure, " *Journal of Pragmatics* 30, pp. 485–496, 1998.

[13] J. E. Arnold, M. Fagnano, M. K. Tanenhaus, "Disfluencies Signal Theee, Um, New Information," *Journal of Psycholinguistic Research,* 32(1), pp. 25-36, 2003.

[14] D. Barr, "Paralinguistic correlates of discourse structure," unpublished, Poster presented at the 42nd Annual Meeting of the Psychonomic Society, Orlando, FL, 2001.

[15] J. Fox Tree, "Listeners' uses of um and uh in speech comprehension," *Memory and Cognition* 29, pp. 320-326, 2001.

[16] H. H. Clark, "Managing problems in speaking," *Speech Communication* 15, 243-250, 1994.

[17] H. H. Clark, J. E. Fox Tree, "Using uh and um in spontaneous speaking, " *Cognition* 84, pp. 73–111, 2002.

[18] E. Shriberg, "Preliminaries to a theory of speech disfluencies," unpublished,. PhD thesis, University of California at Berkeley, 2001.

[19] E. Leeuw, "Hesitation markers in English, German, and Dutch," *Journal of Germanic Linguistics* 19(2), pp.85-114, 2007.

[20] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, N. Hagita, "How quickly should communication robots respond?", HRI Proceedings, 153-160, 2008.

[21] A. Gravano, "Turn-taking and affirmative cue words in task-oriented Dialogue," unpublished, PhD thesis, Columbia University, 2009.

[22] A. Gravano, and J. Hirschberg, "Turn-taking cues in task-oriented dialogue", *Computer Speech and Language*, vol. 25, pp. 601–634, 2011.

[23] Š. Beňuš, A. Gravano, J. Hirschberg, "Prosody of backchannels in American English," Proceedings of 16th International Congress of Phonetic Sciences, 2007.

[24] Š. Beňuš, "Prosodic forms and pragmatic meanings: the case of the discourse marker '*no*' in Slovak," Proceedings of 3rd CogInfoCom conference, 2012.

[25] http://htk.eng.cam.ac.uk/