# Very short utterances in conversation

*Jens Edlund[1], Mattias Heldner[1], Samer Al Moubayed[1], Agustín Gravano[2] and Julia Hirschberg[3]*
[1] *KTH Speech, Music and Hearing*
[2] *Computer Science Department, University of Buenos Aires*
[3] *Department of Computer Science, Columbia University*

## *Abstract*

*Faced with the difficulties of finding an operationalized definition of* backchannels*, we have previously proposed an intermediate, auxiliary unit – the* very short utterance *(VSU) – which is defined operationally and is automatically extractable from recorded or ongoing dialogues. Here, we extend that work in the following ways: (1) we test the extent to which the VSU/NONVSU distinction corresponds to backchannels/non-backchannels in a different data set that is manually annotated for backchannels – the Columbia Games Corpus; (2) we examine to the extent to which VSUs capture other short utterances with a vocabulary similar to backchannels; (3) we propose a VSU method for better managing turn-taking and barge-ins in spoken dialogue systems based on detection of backchannels; and (4) we attempt to detect backchannels with better precision by training a backchannel classifier using durations and inter-speaker relative loudness differences as features. The results show that VSUs indeed capture a large proportion of backchannels – large enough that VSUs can be used to improve spoken dialogue system turntaking; and that building a reliable backchannel classifier working in real time is feasible.*

## Introduction

A large number of vocalizations in everyday conversation are traditionally not regarded as part of the information exchange. Examples include confirmations such as *yeah* and *ok* as well as traditionally non-lexical items, such as *uh-huh*, *um*, and *hmm*. Vocalizations like these have been grouped into different categories and given different names: for example *backchannels* (i.e. back-channel activity, Yngve, 1970), *continuers* (Schegloff, 1982), *feedback* and *grunts*, and attempts at formalizing their function and meaning have been made (e.g. Ward, 2004). The working definitions of these overlapping concepts, however, are imprecise, and different labeling schemes treat them quite differently. The schemes are also often complex. Faced with these difficulties and inspired by others, for example Shriberg et al. (1998), we previously proposed an intermediate, auxiliary unit – the *very short utterance* (VSU) – which is defined operationally and is automatically extractable from recorded or ongoing dialogues (Edlund, Heldner, & Pelcé, 2009). VSUs are intended to capture a large proportion of the interactional dialogue phenomena commonly referred to as backchannels, feedback, continuers, *inter alia*, at zero manual effort.

### VSUs and backchannels

The data we used for our first examination of VSUs, however, was not annotated for backchannels, and automatically identified VSUs were instead compared to annotation for degree of informational content of the same utterances, under the assumption that utterances with low informational content would be representative for backchannels. The first contribution of this paper is to report the extent to which the distinction of VSU/NONVSU, as defined in (Edlund, et al., 2009) captures the distinction between backchannels and non-backchannels as annotated in the Columbia Games Corpus. We also include a more fine-grained analysis of the non-backchannels captured by the VSUs.

Although VSUs may be a useful compromise when no manual annotation is available, we would ideally like to be able to do without them and detect backchannels and other feedback directly. As a first step towards this, we also train a classifier of backchannels on duration and inter-speaker relative loudness and report the preliminary results.

### VSUs and spoken dialogue systems

In spite of the difficulties involved in defining backchannels, there is little controversy surrounding the utility of modeling them. They behave differently from other utterances, and so are interesting both for models of human conversation and for spoken dialogue systems aiming at more human-like behavior. Commonly reported characteristics include the fact that they can be spoken in overlap without disrupting the original speaker (hence the term 'backchannel'). For a spoken dialogue system designer to build systems that encourage users to talk to a system as they would to another human, this phenomenon needs to be managed so that such a system can receive continuous feedback from its users, often in the form of backchannels. As most systems today deal with any user vocalization occurring during system speech as if it were a barge-in, causing the system to stop speaking immediately, the effects of such feedback to current systems would be peculiar and unwanted. The second contribution of this paper is to propose a method to better deal with turntaking in spoken dialogue systems by continuously detecting VSUs and to quantify the potential gain of using such a method.

## Method

### Columbia Games Corpus

The data used in this work is drawn from the Columbia Games Corpus (CGC), a collection of spontaneous task-oriented dialogues by native speakers of Standard American English, and its associated annotations. This corpus contains recordings made using close-talking microphones, with speakers recorded on separate channels, 16 bit/48 kHz, in a sound-proof booth. Speakers were asked to play two types of collaborative computer games that required verbal communication. The speakers did not have eye contact. There were 13 subjects (7 males and 6 females) and they formed 12 different speaker pairs. Eleven of the subjects spoke with two different partners in two separate sessions. The recording sessions lasted on average 45 minutes, and the total duration of the corpus is 9 hours 8 minutes.

The corpus has been orthographically transcribed and manually annotated for a number of phenomena. For the present study, we have used the labeling of single *affirmative cue words* (i.e. lexical items potentially indicating agreement such as *alright, gotcha, huh, mm-hm, okay, right, uh-huh, yeah, yep, yes, yup*) with their communicative function, by three trained annotators, and the labeling of turn-exchanges, by two trained annotators. The function labels for affirmative cue words are *backchannel*, *affirmation/agreement, cue phrase beginning discourse segment, cue phrase ending discourse segment, pivot beginning* and *pivot ending*. Turn exchanges were labeled by first identifying *Interpausal Units* (IPUS), maximal sequences of words surrounded by silence longer than 50 ms (cf. talkspurts in Brady, 1968). A turn was defined as a maximal sequence of IPUS from a single speaker, so that between any two adjacent IPUS there is no speech from the interlocutor (cf. talkspurts in Norwine & Murphy, 1938).

All turn transitions in the corpus were classified using a labeling scheme adapted from (Beattie, 1982) that identifies, inter alia, *smooth switches* (S) — transitions from speaker A to speaker B such that (i) A manages to complete her utterance, and (ii) no overlapping speech occurs between the two conversational turns; *pause interruptions* (PI), defined as cases similar to smooth switches except that A does ***not*** complete her utterance; and *backchannels* (BC), defined as an utterance produced a "response to another speaker's utterance that indicates only *I'm still here / I hear you and please continue*", with no attempt to take the turn. Speech from A following backchannels from B was labeled separately as X2[1].

### Data

For the present study, we used the annotations of turn transitions in silences in the Columbia Games Corpus. We contrasted backchannels with a collapsed non-backchannel category including smooth switches, pause interruptions and utterances following backchannels (S+PI+X2). In addition, we contrasted backchannels with a collapsed category including all other single affirmative cue words (AFFCUE). The backchannel category (BC) in both comparisons was identical, while the other discourse functions of affirmative cue words comprised a subset of the smooth switches plus pause interruptions category.

[1] http://www.cs.columbia.edu/speech/games-corpus/ has further details and annotation manuals.

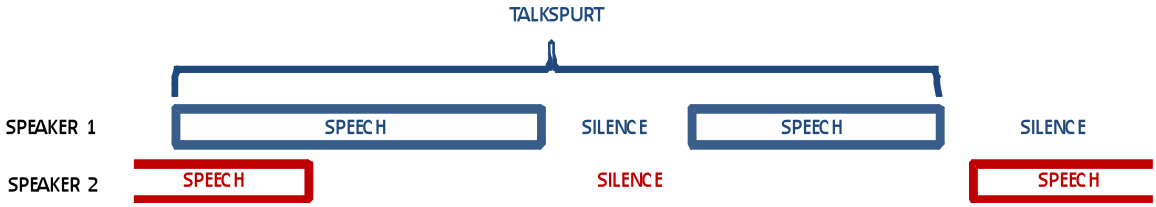


Figure 1. Schematic illustration of a talkspurt as used in the current work.

### Talkspurt durations

We are interested in how long a speaker of a backchannel or a non-backchannel goes on speaking, on average, until the other speaker takes the turn. For this, we need durations of the talkspurt defined by Norwine & Murphy: "A *talkspurt* is speech by one party, including her pauses, which is preceded and followed, with or without intervening pauses, by speech of the other party perceptible to the one producing the talkspurt" (Norwine & Murphy, 1938), or a *turn* in CGC. We note that this definition differs from that used by Brady (1968), in which a talkspurt is a sequence of speech activity flanked by silences in one speaker's channel. Brady's definition has been used by ourselves in previous work (e.g. Edlund & Heldner, 2005; Laskowski, Edlund, & Heldner, 2008), but Norwine & Murphy's concept is better suited for our current purposes, and we adopt their definition in what follows.

### Identifying VSUs

The objective of automatically defining VSUs draws on the observation that backchannels are limited in duration and quiet. Thus, we extracted DURATION from start to finish (see Figure 1) for each talkspurt. A talkspurt was classified as a VSU if the talkspurt's DURATION was shorter than a given threshold. In addition, we extracted loudness differences across turn exchanges using the method and frequency weighting proposed by the ITU (International Telecommunication Union, 2006). In (Edlund, et al., 2009) we also used voicing ratio which helped filter out mistakes made by the automatic speech detector. As the CGC data is manually annotated, we left the voicing ratio parameter out for simplicity.

### Backchannel classifier training

For training, the K-Nearest Neighbors (K-NN) method was used. K-NN is a non-parametric non-linear classifier which does not build a model for training data, but builds a local model for each test sample using that sample's neighborhood. The study in (Atkeson, Moore, & Schaal, 1997) gives a good overview of the method. Initially, the K-nearest neighbors to the test sample are collected using Euclidian distance on the features. Then, a weighted average voting of these neighbors decides which class the test sample belongs to. In our case we use a binary classification of BC/NONBC, and the method would give a number between 0 and 1, which can be taken as the probability of the class 0 or 1. Classifiers were trained using duration only (DUR) as well as duration and inter-speaker loudness difference (DUR+LOUDDIFF).

## Results

Table 1 shows the confusion matrix for VSU/NONVSU versus the manual annotation of BC/NONBC using the same threshold for VSU as in our previous study: 1 s. We note in particular that all BCs in the material are also VSUs with this threshold. Figure 2 shows the underlying data – the histograms over the durations of BC and NONBC. Table 2 shows a cross tabulation of VSU/NONVSU (again using a 1 s threshold) versus the manual annotation of BC/AFFCUE/OTHER, and the underlying data for this three-way split – the histograms over the durations of BC, AFFCUE and NONBC appear in Figure 3.

Table 1. Confusion matrix for VSU/NONVSU versus the manual annotation of BC/NONBC.

| | VSU | NONVSU | TOTAL |
|---|---|---|---|
| BC | 553 | 0 | 553 |
| NONBC | 1208 | 2600 | 3808 |
| TOTAL | 1761 | 2600 | 4361 |

Table 2. Cross tabulation of VSU/NONVSU versus a manual annotation of BCS versus AFFCUE and OTHER.

| | VSU | NONVSU | TOTAL |
|---|---|---|---|
| BC | 553 | 0 | 553 |
| AFFCUE | 699 | 768 | 1467 |
| OTHER | 509 | 1832 | 2341 |
| TOTAL | 1761 | 2600 | 4361 |

We note that 31% of the VSUs are backchannels and 40% are AFFCUES. Inspection of the remaining 29% of VSUs labeled as OTHER revealed that a large proportion of them also have feedback functions. The 25 most frequent tokens are *mm*, *no*, *oh*, *got it*, *um*, *oh okay*, *hm*, *I'm gonna pass*, *uu*, *I have to pass*, *cool*, *great*, *nope*, *and*, *sure*, *that's it*, *and then*, *don't have it*, *exactly*, *I don't have that*, *oh right*, *oh yeah*, *so*, and *sorry*. These comprise about one third of all OTHER VSU tokens.

Moving on to the BC/NONBC classifiers, we observe that the classifiers trained on duration only (DUR) and duration plus relative loudness (DUR+LOUDDIFF) showed similar performance, with a slight advantage for the combination of the two. Figure 4 shows the ROC curves for the DUR classifier and the DUR+LOUDDIFF classifier.

The DUR+LOUDDIFF classifier was applied using a 77% training set, and 23% test set size, resulting in 1000 test samples (890 NONBC, and 110 BC). Using an optimized threshold of 0.0995, the overall accuracy of the system resulted in 73% correct classification. Table 3 presents the confusion matrix between the two classes on the test set.

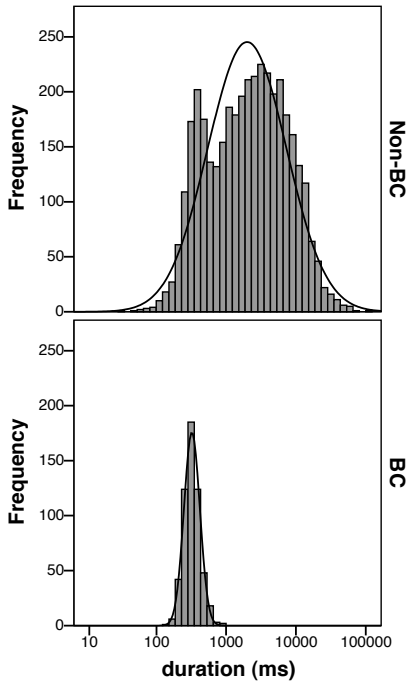


Figure 2. Histograms over durations in ms of manually annotated NONBCS (top) and BCS (bottom) in CGC.

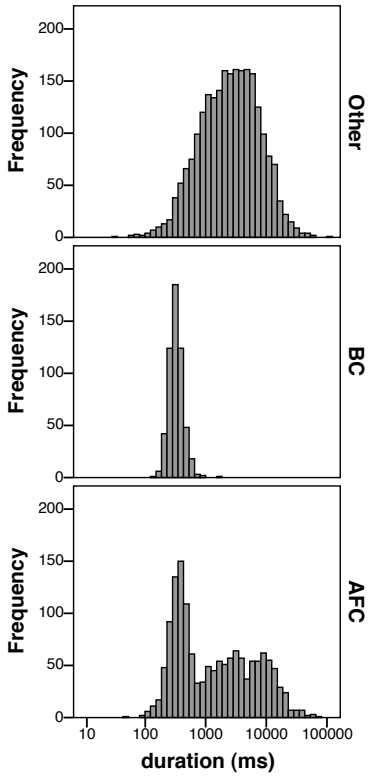


Figure 3. Histograms over durations in ms of manually annotated AFFCUE (bottom), BC (middle) and OTHER (top) in CGC.

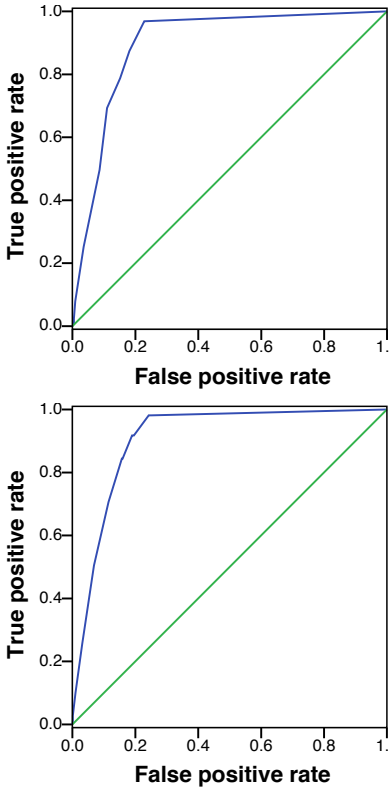


Figure 4. ROC curves for the DUR classifier (top) and the DUR+LOUDDIFF (bottom). The areas under the curves are 0.896 and 0.908, respectively.

Table 3. Confusion matrix between the two classes on the manually annotated test set.

| | | PREDICTED | |
|---|---|---|---|
| | | NONBC | BC |
| TRUE | NONBC | 70.7865 | 1.8182 |
| | BC | 29.2135 | 98.1818 |

# Discussion

The results in Tables 1 and 2 show that the 1 s duration threshold we previously used to automatically identify VSUs indeed captures a large portion – all, in fact – of the manually annotated backchannels in the Columbia Games Corpus. The duration distributions of BCs and NONBCS, respectively, in Figure 2 suggests that more precise discrimination of backchannels is possible in this material using duration alone, as the majority of backchannels are considerably shorter than 1 s. A shorter threshold would eliminate much of the NONBCS identified as VSUs.

If we keep the 1 s threshold, we note that amongst the talkspurts that are VSUs, yet are not backchannels, many belong to the group of other affirmative cue words, so that the majority (71%) of talkspurts identified as VSUs are either BCs or AFFCUE. Of the remaining 29%, the 25 most frequent words can all be ascribed feedback functions. The 1 s VSUs, in other words, capture almost exclusively short feedback utterances, and using a lower threshold will increase the proportion of manually annotated backchannels.

Although VSUs appear to be a good approximation of backchannels (and other feedback, depending on the duration threshold), we would prefer a classifier that could, in real-time and immediately at the beginning of a talkspurt, identify backchannels. The BC/NONBC classifier is a first step towards this, and its performance suggests that using duration and inter-speaker relative loudness only to train a classifier seems viable. The ROC curves in Figure 4 are promising, and the result of running the combined classifier on unseen test data suggest that BCs can be detected reliably using these features.

Finally, from a speech technology perspective, the histogram in Figure 3 is encouraging. We see that the vast majority of backchannels are shorter than 500 ms, which makes the following strategy for barge-ins tractable:

When user speech is detected during a system utterance, do the following:

- Go on speaking for 300-500 ms.
- If the user has stopped speaking after 300-500 ms has passed, the vocalization was likely a backchannel, so just go on.
- If, on the other hand, the user is still speaking after 300-500 ms, the vocalization is highly unlikely to be a backchannel, so consider stopping (for a polite system) or raising the system's voice (for urgent messages or for an impolite system).

If we allow for 200 ms to detect silence, approximating the detection thresholds for humans (cf. Izdebski & Shipp, 1978; Shipp, Izdebski, & Morrissey, 1984) as well as for many voice activity detectors (e.g. VADER[2]),

[2] See the CMU Sphinx Speech Recognition Toolkit: http://cmusphinx.sourceforge.net/

the system should be able to make an informed decision at the expense of occasional latencies. On the the Columbia Games Corpus data, such a system would never mistake a backchannel for a barge-in, at the expense of 500-700 ms response delays *occurring only when the user barges in*. This delay corresponds roughly to two or three syllables of speech.

# Conclusions

We have shown that the VSU – our previously proposed automatically extractable auxiliary unit – does indeed capture, with zero manual labor, a large proportion of talkspurts annotated as backchannels in the Columbia Games Corpus. Furthermore, a large proportion of those VSUs that are not backchannels are instead different forms of affirmative cue words and other types of feedback which from a dialogue system point of view may be treated in a similar manner. We have trained a BC/NONBC classifier on duration and inter-speaker relative loudness, and found that it finds backchannels with high accuracy and that adding the relative loudness may yield a performance improvement, which is consistent with the claim that backchannels are quiet. This is a first step towards eliminating, at least in part, the intermediate VSU classification.

We have also suggested a method which utilizes the shortness of backchannels to avoid having a barge-in sensitive spoken dialogue system halt abruptly at each backchannel. The cost of this method is acceptable at a latency of some 500-700 ms, applied only where the user speaks at the same time as the system.

We conclude by noting that these findings all point to backchannels being unobtrusive and acoustically not very prominent, and that they are all consistent with descriptions of them as being relatively brief, soft and quiet.

# Acknowledgements

This research was carried out at the Department of Computer Science, Columbia University, New York. Funding was provided by the Swedish Research Council (VR) project 2009-1766 *The Rhythm of Conversation*, and by NSF IIS-0307905.

# References

Atkeson, C, Moore, A, & Schaal, S (1997). Locally weighted learning. *Artificial intelligence review,* 11: 11–73.

Beattie, G W (1982). Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. *Semiotica,* 39: 93-114.

Brady, P T (1968). A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal,* 47: 73-91.

Edlund, J, & Heldner, M (2005). Exploring prosody in interaction control. *Phonetica,* 62: 215-226.

Edlund, J, Heldner, M, & Pelcé, A (2009). Prosodic features of very short utterances in dialogue. In: M Vainio, R Aulanko & O Aaltonen, eds, *Nordic Prosody: Proceedings of the Xth Conference, Helsinki 2008*. Frankfurt am Main: Peter Lang, 57-68.

International Telecommunication Union. (2006). Recommendation ITU-R BS.1770-1: Algorithms to measure audio programme loudness and true-peak audio level.

Izdebski, K, & Shipp, T (1978). Minimal reaction times for phonatory initiation. *Journal of Speech and Hearing Research,* 21: 638-651.

Laskowski, K, Edlund, J, & Heldner, M (2008). An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems. In *Proceedings ICASSP 2008*. Las Vegas, NV, USA, 5041-5044.

Norwine, A C, & Murphy, O J (1938). Characteristic time intervals in telephonic conversation. *The Bell System Technical Journal,* 17: 281-291.

Schegloff, E (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In: D Tannen, ed, *Analyzing Discourse: Text and Talk*. Washington, D.C., USA: Georgetown University Press, 71-93.

Shipp, T, Izdebski, K, & Morrissey, P (1984). Physiologic stages of vocal reaction time. *Journal of Speech and Hearing Research,* 27: 173-178.

Shriberg, E, Bates, R, Stolcke, A, Taylor, P, Jurafsky, D, Ries, K, et al. (1998). Can prosody aid in the automatic classification of dialog acts in conversational speech. *Language and Speech,* 41: 439-487.

Ward, N (2004). Pragmatic functions of prosodic features in non-lexical utterances. In *Proceedings of Speech Prosody 2004*. Nara, Japan, 325-328.

Yngve, V H (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting Chicago Linguistic Society*. Chicago, IL, USA: Chicago Linguistic Society, 567-578.