The Prosody of Swedish Conversational Grunts

D. Neiberg, J. Gustafson

Centre for Speech Technology (CTT), TMH, CSC, KTH, Stockholm, Sweden

[neiberg,jocke]@speech.kth.se

Abstract

This paper explores conversational grunts in a face-to-face setting. The study investigates the prosody and turn-taking effect of fillers and feedback tokens that has been annotated for attitudes. The grunts were selected from the DEAL corpus and automatically annotated for their turn taking effect. A novel suprasegmental prosodic signal representation and contextual timing features are used for classification and visualization. Classification results using linear discriminant analysis, show that turn-initial feedback tokens lose some of their attitude-signaling prosodic cues compared to non-overlapping continuer feedback tokens. Turn taking effects can be predicted well over chance level, except Simultaneous Starts. However, feedback tokens before places where both speakers take the turn were more similar to feedback continuers than to turn initial feedback tokens.
Index Terms: prosody, fillers, feedback, suprasegmental, conversational grunts

1. Introduction

Conversation is the most common use of speech. Any automatic dialog system, pretending to mimic a human, must be able to successfully detect typical sounds and meanings of spontaneous conversational speech. This task may be implemented as automatic transcription or classification of Dialog Acts (DAs). This can be done on the lexical level, on the prosodic level [1][2], or on both [3][4].

The present study investigates conversation grunts, that are either words like "okey" and "yes" or non-lexical tokens like "mhm" and "eh" [5]. These conversational tokens can be roughly divided into those that are interjected into one's own speech (fillers) and those that are interjected into the interlocutor's speech (feedback). The filler "um" has been found to be the 6th most frequent item in the Switchboard Corpus [6] and back-channels has been found to account for 19 % of the dialogue acts in a subset of the same corpus [5]. Feedback tokens are usually divided into yes/no answers, back-channels and acknowledgments. This study also investigates the prosodic cues to the perceived attitude of the feedback tokens. In our corpus the following feedback attitudes have been manually annotated: dis-preference, news receiving and general feedback. Depending on the context and prosodic realization, the same feedback token can have quite different meanings. This means that in order to automatically assign meaning to conversational grunts it is essential to take into account their context and model their prosodic realizations. Finally, both fillers and feedback tokens have been annotated for their turn-taking effect (i.e. who speaks after a produced grunt): Other Speaker, Same Speaker or Simultaneous Starts. Our main hypothesis is that conversational grunts are carriers of prosodic information, and this study shows how their prosodic realization signal attitude and turn taking intention.

For classification and intuitive visualization of feedback and fillers, we use a supra-segmental prosodic signal representation based on Time Varying Constant-Q Cepstral Coefficients (TVCQCC) introduced in [7]. The TVCQCC are suitable for machine learning of varying length segments, and visualization of common properties shared by multiple segments. It allows for direct modeling of the F0 region of spectra with less complexity than a pitch tracker. The contribution of the end of the interlocutor left context for predicting the turn taking effect [8] is used to boost classification in this study. In addition, we examine the contribution of a variant of contextual timing features, which has been shown to be useful in DA recognition [4].

2. The DEAL corpus

This study uses data from the DEAL corpus [9]. It consists of dialog data recorded as an informal, human-human, face-to-face conversation. The data collection was made with 6 subjects (4 male and 2 female), 2 posing as shop keepers and 4 as potential buyers. Each customer interacted with the same shop-keeper twice, in two different scenarios. The customers were given a task: to buy items at a flea market at the best possible price.

All dialogs in the DEAL corpus were transcribed orthographically including non-lexical entities such as laughter and audible breathing. Filled pauses, repetitions, corrections, restarts and cue phrases were labeled manually. The corpus is rich in fillers and feedback tokens. The feedbacks were generally single words or non-lexical tokens and appeared in similar dialog contexts (i.e. as responses to assertions). The feedbacks are labeled according to their perceived attitudes; news receiving, dis-preference or general feedback. For this study, only the tokens which resemble the list of conversational grunts found in [5] were used. These units have also been referred to as minimal listener response tokens [10]. Apart from non-word vocalizations these also include some feedback words where the meaning is heavily dependent on the context and prosody. These are broadly speaking variations of "okey", "yes", and "no", which in Swedish are translated to "okej", "ja" and, "nej". In spontaneous dialogs they often occur in reduced forms like "a" and "nä" as well as in in-between versions like "njo" and "jaä".

3. A Suprasegmental Fundamental Frequency Representation

Instead of using prosodic features derived from pitch-tracker, we use a special case of Time Varying Constant-Q Cepstral Coefficients (TVCQCC). The filter-bank is based on the Constant-Q transform with a corresponding Q factor of $1/(2^{1/12}-1)$ or 16.8 which corresponds to the 12 semitones per octave in a musical scale. Here the filter-bank spans a total of 81 bins between 60 Hz and 6458 Hz, which is below the Nyquist frequency. Compared to Short-time Fourier Transform (STFT),

Copyright © 2010 ISCA

the constant-Q transform has optimal temporal-spectral resolution for all filters, which means there is no need to optimize the analysis window length for different applications. A standard frame shift rate of 100 Hz is used.

The per token average F0 is found by first summing harmonics for each filter in the semitone scale per frame. The maximum number of harmonics to sum over is 12 because beyond that consecutive harmonics would fall under the same bin. In order to give a reasonable resolution for high frequencies only the first 8 harmonics were used in this study An approximation of F0 to noise separation is used which classifies all frequencies with amplitudes below 10 dB from the highest amplitude frequency component as noise. So any local maxima above this threshold occurring in the output of the filter-bank are considered as tones, which means that the summing starts at the first index containing non-noise. The per-frame estimated F0 is then found by the semitone corresponding to maximum of the harmonic summation. Then the average F0 is found by a power amplitude weighted average of the per-frame estimated F0s. This is motivated by a previous study that found frequencies at higher intensity levels to be more salient [11], and since it removes the need for voicing decision. The range is kept within 8 semitones from the mean frequency, which leads to the assumption of a maximum F0 variation of 17 semitones. However, this choice reduces the influence of the first overtone which is located at a distance of 1 octave.

We propose a supra-segmental parametrization for each token instead of a frame based representation. The proposed method provides a suitable way of integrating the information available in a token into a matrix of fixed size. First the log time spectrum $lX(k,n)$ for every frame n and every filter k is obtained. Then, the TVCQCC are calculated by applying a 2 dimensional discrete cosine transform (2D-DCT), as follows. For $1 \leq p \leq P$ and $1 \leq q \leq Q$, (where P and Q are the number of coefficients in the frequency and time dimension respectively), the TVCQCC are

$$T(q,p) = \sum_{n=1}^{N}\sum_{k=1}^{K} \frac{lX(k,n)}{N} * cos\left(\frac{\pi(k-\frac{1}{2})(q-1)}{K}\right) * cos\left(\frac{\pi(n-\frac{1}{2})(p-1)}{N}\right) \quad (1)$$

The axis of T along q is called the 'quefrency' and has a time dimension. The axis along p is the frequency of quefrency, which here is referred to as 'meti' (following the convention of swapping syllables), and has frequency dimension. This dimension space is also known as cepstrum modulation spectrum. It should be noted that the 2D-DCT has been modified so that this representation is length invariant, which means that the parameters are not affected by stretching or compression in time.

4. Experiments

4.1. The turn-taking effect of grunts

In the investigation presented in this paper three types of turn-taking effects for both fillers and feedback tokens are considered: Same Speaker, Other Speaker and Simultaneous Start, illustrated in Figure 1. These definitions always use the speaker who uttered the filler or feedback as reference. Same Speaker means a non-overlapping floor taking for the reference speaker, while the Other Speaker condition implies the floor is immediately given back. Simultaneous Start is when both speakers starts within a 300 ms time-frame from each other. The turn

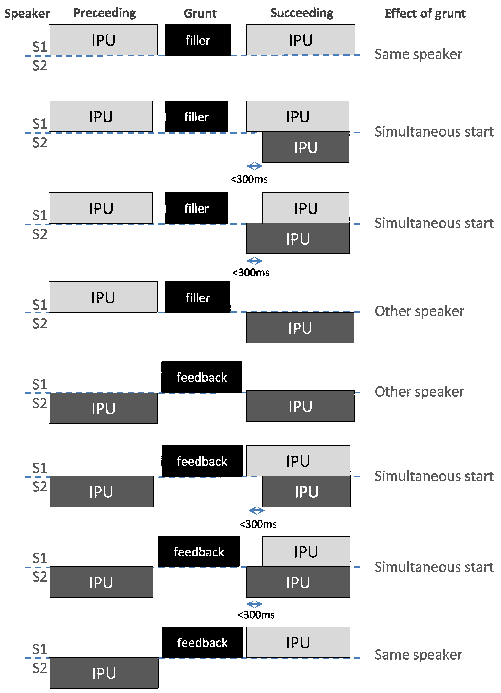

Figure 1: *Proposed turn-taking effect of grunts defined with Inter Pausal Units (IPUs).*

switches which had laughter or hawks in the intermediate context were all removed. Tokens shorter than 40 ms and tokens with an estimated mean F0 above 600 Hz were also removed, but they were very few. The occurrence's in data of these conditions for fillers and the three feedback attitudes are shown in Table 1.

4.2. Automatic Classification

This section main objectives are: 1) determine if the proposed prosodic features can discriminate between fillers and feedback attitude. 2) determine if their turn-taking effect can be predicted 3) determine the contribution of timing features and end of interlocutor left context. Thus, the following tasks were considered:

Task 1 Fillers vs. Feedback where all attitudes are merged

Task 2a Feedback attitudes (3 classes)

Task 2b Same as 2a but only for the Other Speaker condition

Task 2c Same as 2a but only for the Same Speaker condition

Task 3a The turn-taking effect of feedbacks (Other Speaker / Same Speaker / Simultaneous Start)

Task 3b Same as 3a, but with Simultaneous Starts removed

Task 3c Same as 3a, but Other Speaker and Simultaneous Start are merged

Task 3d Same as 3a, but Same Speaker and Simultaneous Start are merged

Fillers turn-taking effect was not addressed since they too seldom caused a turn shift, resulting in too little training data.

For classification, $Q = 8$ quefrency and $P = 8$ meti coefficients are used which gives a reasonable resolution in frequency and time. The TVCQCC matrices are converted into vectors by stacking the rows after each other, then the estimated mean F0 and token length was added. Classification of the resulting vectors is done by using linear discriminant analysis with diagonal

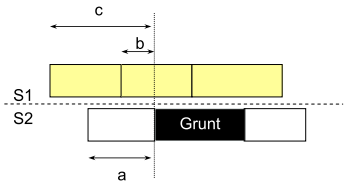

Figure 2: *First three timing features as durations a,b and c.*

covariance to ensure robustness, which is critical when the feature dimension is in the same magnitude as the number of training samples. The LDA priors are set to equal to avoid a bias toward speaker or domain dependent behavior. The choice for evaluation is 8 fold cross-validation on the dialog level.

The timing feature vector has four dimensions, were the first three are shown as durations of silence or tokens in Figure 2. The fourth is the duration of the gap or overlap (where duration is assigned a negative value). If the grunt token is in complete overlap, which may be the case for some back-channels, then the negative duration of the grunt token is used as a back-off value. All tasks are evaluated with or without timing features as auxiliary information.

The left-context is the immediate non-silence part of the interlocutor parametrized by TVCQCC. If the interlocutor is silent during the segment which partly overlaps the grunt tokens' start in time, then the preceding token is used as context. If the interlocutor speaks during the partly overlapped segment, then only the part until the speaker change is used, but if there exist a preceding token, it is also used as left context. In the case where the interlocutor left context was included, the TVCQCC matrix was simply calculated for the context and added to the original feature vector which doubles the feature dimension. Tasks 3a-d are tested with or without the left context of the interlocutor, while the left-context has no relevance for Task 1 and 2 and just introduced noise which degraded the classifier performance.

The results are shown in Table 2. A brief look shows an overall boost if timing features are added while the left-context boosts turn-taking tasks. Average recall, defined as the average recall rate of all classes is reported and the corresponding random guess.

Table 1: *Occurrences of tokens relevant for the study.*

Type	Other Spk.	Same Spk.	Sim.Start
General F.b.	177	92	22
Dispreference F.b.	30	60	3
News receiving F.b.	31	33	9
Filler	20	148	12

4.3. Plotting Prototypical Spectrograms

In order to facilitate visualization, a way of plotting *prototypical spectrograms* was introduced in [7]. Many classifiers, such as LDA or Naïve Bayes, use a Gaussian distribution as underlying parametrization. If we want to see what the classifier relies on, then the multivariate mean value is the natural starting point. Thus, the basic idea is to take the average TVCQCC of all instances for each class, followed by inverse 2D cosine transformation. Displaying the essence using the average instead of the accumulation of frequencies as in [12] has thus different purposes. Figure 3 shows *prototypical spectrograms* with power

Table 2: *Results measured in average recall for all tasks. Standard deviation is between 1.0-1.3 for all ratios.*

Task	Timing	No Context	Left Context	Random
1	No	79.8	-	50
1	Yes	82.1	-	50
2a	No	47.8	-	33
2a	Yes	48.6	-	33
2b	No	49.1	-	33
2b	Yes	50.4	-	33
2c	No	42.9	-	33
2c	Yes	45.1	-	33
3a	No	39.1	41.2	33
3a	Yes	39.7	43.0	33
3b	No	63.9	65.3	50
3b	Yes	66.0	67.2	50
3c	No	64.1	64.9	50
3c	Yes	64.8	66.7	50
3d	No	62.0	64.7	50
3d	Yes	62.8	64.8	50

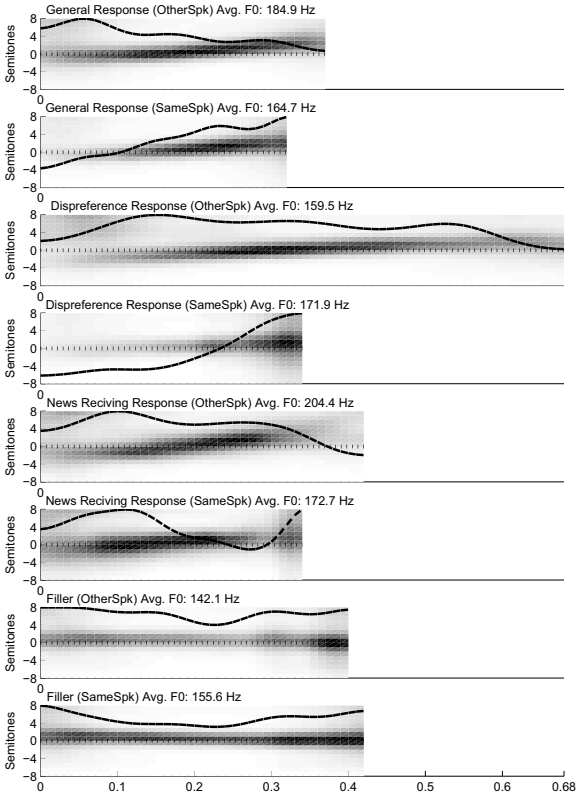

Figure 3: *Prototypical spectrograms with power amplitude scale for fillers and feedbacks divided into Other Speaker and Same Speaker turn-taking conditions. The y-axis shows semitones relative to the average F0 of the token, and the x-axis shows time in seconds. The dashed lines are energy curves calculated as the sum of all filters from the full frequency range.*

amplitude scale for fillers and feedbacks subdivided into their turn-taking effect. Simultaneous Starts are not shown due to space constraints. Each spectrogram is stretched to the average duration to further show the characteristics of each type and the average F0 is given in Hertz. In addition, the energy calculated

as the sum of all filters of the full frequency range is plotted in the same graphs, but scaled to the same maximum for all categories.

5. Results and discussion

The spectrograms in Figure 3 show notable differences. First, feedback tokens have a raise in F0 while fillers are flat or have a slight drop. Fillers has also a lower average F0 compared to feedbacks. Among the feedback attitudes for the Other Speaker condition, dis-preference is longer than both News Receiving and general feedback. News receiving has the highest average F0 and a strong rise while Dis-preference has the lowest average F0 and a weaker rise. These findings closely follows the results for Bad News (Dis-preference) vs Good news (News receiving) found in English response tokens (with a slightly different definition than Ward) [13]. Feedbacks have shorter duration and a final energy rise in turn initial position compared to feedback tokens that act as continuers. The intensity curves for feedback tokens that give back the turn to the interlocutor indicate that general feedbacks seems to be monosyllabic, news receiving feedback tend to be bisyllabic while dis-preference feedback is harder to characterize. While the turn-taking effect of filled pauses was not examined in any of the classification tasks, Fillers that are followed a speaker shift seem to have a lower initial amplitude of voicing that those that are followed by more speech from the same speaker.

The recall rates for Task 1 as well as the spectrograms indicate a clear difference between filled pauses and feedback tokens. Task 2 shows that it is possible to discriminative between the three feedback attitudes, with better results in cases where the feedback tokens gives back the turn to the interlocutor. This is not that surprising since turn initial feedback tokens simply functions as floor-grabbers, where the attitudinal content can be communicated later in the turn. The fact that the human annotator could discriminate between feedback attitudes for the turn initial feedback tokens may be explained by the fact that they had access to the previous and following turns as well as the lexical and prosodic realization of the feedback token under investigation. Task 3a recall rate shows an above chance level for the three proposed turn-taking effects, but a closer look at the confusion matrix showed a recall rate below chance for Simultaneous Starts. Removing these, as showed in Task 3b gives well above chance average recall rate, and the left context gives additional boost. The boost from the left context may come from the difference between back-channel eliciting cues, which may boost Other Speaker decisions, as opposed to phrase-final intonation which may boost Same Speaker decisions. Comparing Task 3c and 3d shows merging Simultaneous Starts and Other Speaker gives the same results as for Task 3b, while merging Simultaneous Starts with Same Speaker condition degrade performance. This indicates Simultaneous Starts are preceded by the same feedback and context realizations as for the Other Speaker condition, which indicates what we see floor stealing attempts that are not prosodically signaled by the feedback producing speaker.

6. Conclusions

In this paper the prosody of conversational grunts has been explored using a novel supra-segmental signal representation for the F0-region. The results show clear discriminative ability between fillers and feedbacks. The proposed signal representation is shown to be able to discriminate between dis-preference, news receiving and general feedback grunts, especially if the feedback producing speaker gives back the floor without non-overlapping speech, otherwise only smaller prosodic differences were found. In addition, the contribution of contextual timing features improve the performance for all tasks. Further, it is possible to predict the turn taking effect well above chance, unless there is a Simultaneous Start. For this latter task, using the left context of the interlocutor boosts performance. These findings may be used to design a more general detector for these social interaction phenomenons.

7. Acknowledgments

This research is carried out at KTH Speech, Music and Hearing. Funding was provided by the Swedish Research Council (VR) project 2009-4291. The authors would like to thank Anna Hjalmarsson for proving the DEAL corpus with annotations for fillers and feedback and for commentating on a draft of this paper. Special thanks goes to Ananthakrishnan Gopal who gave significant contributions in developing TVCQCC.

8. References

[1] Laskowski, K. and Shriberg, E., "Comparing the contributions of context and prosody in text-independent dialog act recognition," in 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2010). Dallas TX, USA, March 2010.

[2] Goto, M., Itou, K., , and Hayamizu, S., "A real-time filled pause detection system for spontaneous speech recognition," in Eurospeech '99, 227–230, September 1999.

[3] Sridhar, V. K. R., Bangalore, S., and Narayanan, S., "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging," Computer Speech and Language, 23(4):407 – 422, 2009.

[4] Gravano, A., Benus, S., Hirschberg, J., Mitchell, S., and Vovsha, I., "Classification of discourse functions of affirmative words in spoken dialogue," in Interspeech, Antwerp, 1613–1616, 2007.

[5] Ward, N., "Issues in the transcription of english conversational grunts," in Proceedings of the 1st SIGdial workshop on Discourse and dialogue, Morristown, NJ, USA, 29–35, Association for Computational Linguistics, 2000.

[6] Ward, N., "The challenge of non-lexical speech sounds," in In International Conference on Spoken Language Processing, 2000.

[7] Neiberg, D., Laukka, P., and Ananthakrishnan, G., "Classification of affective speech using normalized time-frequency cepstra," in Prosody 2010, May 2010.

[8] Duncan, S., "Some signals and rules for taking speaking turns in conversations," Journal of Personality and Social Psychology, 23:283–292, 1972.

[9] Hjalmarsson, A., "Speaking without knowing what to say... or when to end," in Proceedings of SIGDial 2008, Columbus, Ohio, USA, jun 2008.

[10] O'Keeffe, A. and Adolphs, S., *Response tokens in British and Irish discourse. Corpus, context and variational pragmatics*, chapter Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages, 69 – 98, Pragmatics & Beyond New Series 178. Benjamins, Amsterdam/Philadelphia, 2008.

[11] Moore, B. C. J., *An Introduction to the Psychology of Hearing*, Academic Press Limited, 3rd edition, 1989.

[12] Edlund, J., Heldner, M., and Pelcé, A., "Prosodic features of very short utterances in dialogue," in *Nordic Prosody - Proceedings of the Xth Conference*, Vainio, M., Aulanko, R., and Aaltonen, O., Eds., Frankfurt am Main, 57 – 68, Peter Lang, Oct 2009.

[13] Freese, J. and Maynard, D. W., "Prosodic features of bad news and good news in conversation," Language in Society, 27(2):195–219, 1998.