Prosodic cues to engagement in non-lexical response tokens in Swedish

Joakim Gustafson and Daniel Neiberg

Department of Speech, Music and Hearing, KTH, Sweden

jocke@speech.kth.se, neiberg@speech.kth.se

Abstract

This paper investigates the prosodic patterns of non-lexical response tokens in a Swedish call-in radio show. The feedback of a professional speaker was investigated to give insight in how to build a simulated active listener that could encourage its users to continue talking. Possible domains for such systems include customer care and second language learning. The prosodic analysis of the non-lexical response tokens showed that the engagement level decreases over time. Prosodic cues to this include change in syllabicity, pitch slope and loudness. We have also investigated prosodic alignment, to see to what extent the active listener mimic the prosody of the callers in his non-lexical response tokens.

Index Terms: listener responses, prosodic cues, turn management, prosodic alignment

1. Introduction

Today's spoken dialogue systems are being considered for areas and applications beyond simple directory inquiries and travel booking – such as social and collaborative applications, education and entertainment. These new areas call for systems to be increasingly human-like in their conversational behavior. In human-human conversations both parties continuously and simultaneously contribute actively and interactively to the conversation. Listeners actively contribute by providing feedback, and they continuously monitor the speaker contributions for cues allowing smooth speaker shifts. Their feedback indicates attention, feelings and understanding, and its purpose is to support the interaction [1]. *Listener responses* can be categorized according to form/function, verbal/non-verbal lexical/non-lexical and turn/backchannel [2]. Studies on listener responses have used terms like *response tokens* [3], *backchannel continuers* [4], *back-channel grunts* [5], *listener vocalizations [6]*, and *feedback morphemes* [7]. For a more extensive inventory of terms related to listener responses see the list in Fujimoto 2007 [2]

According to Gardner different response tokens have different functions [3]. The bisyllabic "mhm" and "uh-huh" often function as *continuers*. This has also been found for Japanese back-channel grunts where multiple syllables indicate a lack of anything to say [5]. Monosyllabic "yeah" and "mm" are usually are used as *acknowledgement* or *assessment*. The pitch contour of the non-lexical "mm" determines its meaning. A falling pitch is associated with completion and segmentation, as well as with unexcitement and low involvement. A fall-rise contour makes the "mm" into a continuer like "mhm". It often occurs after pragmatically incomplete turns, and it encourages the other speaker to continue talking. A rise-fall contour is associated with heightened involvement and interest, as well as with assessment of the previous turn.

According to Ward the important prosodic features of conversational grunts are: loudness, height and slope of pitch, duration, syllabification, duration and abruptness of the ending [5]. Most of these features were used in a study on the prosody of acknowledgements and backchannels in task oriented dialogues [8]. They found that backchannels in general higher in pitch and intensity than acknowledgements.

2. Background

In order to develop systems that can achieve the responsiveness and flexibility found in human-human interaction, it is essential that they process information incrementally and continuously rather than in turn sized chunks [9]. Early speech synthesizers were reading machines – machines that read written text out loud. From then on, the fact that read speech is very different than speech in interaction has had little impact on speech synthesizers. Conversational grunts, audible breathing and self-corrections are abundant in conversational speech. Being less common in read speech, they are systematically removed in speech synthesis regardless of synthesis method; the rationale being that they do not carry propositional content, and today's synthesizers are optimized to transmit the propositional content of a message. However, regardless of propositional content, they are of immense importance for the interactional aspects of conversation, and without them, we are left strangely lacking in interactional skills.

Our group has a long-standing interest in human conversational behavior and a special interest in its mimicry and evaluation in spoken dialogue systems [10]. We have in a previous study examined the benefit of adding conversational grunts in a commercial call routing system system [11]. This study showed that the addition of backchannel continuers, like "mhm" made the customers more talkative in their problem descriptions, which facilitated a more fine-grained call routing.

Encouraged by this result we have recently initiated a three-year research project that aims at adding human interactional verbal behavior in speech synthesis. The first phase of the project deals with conversational grunts by: (1) annotating instances of them in corpora of human-human conversations, (2) synthesizing the missing tokens using several methods, and (3) evaluating the results in a series of experiments comparing synthesized versions with the originals as well as evaluating their perceived meaning and function. This paper investigates the prosody of Swedish non-lexical response tokens.

3. The active listener database

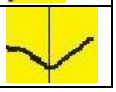
In the current study we have analysed response tokens in a corpus of 73 calls to a Swedish phone-in radio program. The program is called *Ring P1*, and it allows members of the public to call in and share their opinions on current affairs. These interactive "letter to the editor"-calls are handled by popular Swedish journalists. We have selected six 45-minute programs hosted by the most experienced moderator (a blind journalist called Täppas Fogelberg). We selected to study his response tokens since he is known for being a good listener. The structure of this kind of dialogue has been investigated in previous studies [e.g. 12]. Phone-in calls are usually initiated in an opening phase, where the callers exchange greetings with the radio host. In the main phase the callers give their opinion about the some urgent topic, during which the radio host either provides encouraging feedback or engage in a discussion about the topic. This phase is typically one to five minutes long (on average about three minutes). In the last phase the radio host firmly ends the call by telling the caller

that they should give room for other callers. This phase is usually initiated by a loud high-pitched "*Du..*" (You..), or by the host promptly switching to a completely different, but less engaging topic like "*So did you watch the Eurovision song contest?*". In this study we have selected the main phases of 73 dialogues - a dialogue corpus of about three hours. During the main phase the callers produced on average 22 inter pausal units (IPUs) that in half of the cases were followed by speaker shifts. In about 80% of these speaker shifts the radio host merely produced short backchannel continuers that encouraged the caller to continue speaking. This means that the radio host mostly acted as an *active listener*.

3.1. Data selection and tagging

Since the recordings of the radio programs are recorded in mono the first step was to manually annotate the speech for speaker, (where overlapped speech was labelled as both). The syllable boundaries of the last three syllables of the caller IPUs were manually assigned. The response tokens were tagged as lexical (e.g. "*ja*", "*ok*") or non-lexical (e.g. "*mhm*" and "*mm*") and as monosyllabic ("*mm*" and "*ja*") or bisyllabic ("*mhm*" and "*jaha*"). In the 73 dialogues there where 174 lexical and 459 non-lexical response tokens, out of which 44% were perceived as bisyllabic. In this study the prosodic patterns of the non-lexical response tokens "*mm*" and "*mhm*" have been investigated. For these response tokens pitch contour, intensity distribution and syllable boundaries were manually labelled. In Table 1 the appearance of the most common prosodic contours are visualized in pitch curves where the line width indicate the intensity, and the vertical line in the bisyllabic fall-rise cases mark the syllable boundaries.

Table 1. *Examples of intensity modulated pitch curves, with pitch movement in rows and intensity distribution in columns.*

	early	**late**	**two peaks**	**even**
fall				
rise				
fall-rise				

All response token were also labelled for engagement level, where passive corresponds to acknowledgement that the radio host is still listening, while active response tokens signal interest and encourages the caller to say something more. The pitch slope of the loudest part of the pitch curves of the bisyllabic tokens correlates closely to the perceived engagement of the feedback: 90% of the bisyllabic response tokens that had a falling pitch on the loudest syllable were perceived as passive, while 80% of the tokens with rising pitch on the prominent syllable were perceived as active. In bisyllabic response tokens with two intensity peaks or even intensity there was a 50/50 split in the engagement ratings. The perceived engagement level varied over time, see Table 2.

Table 2. *Percentage of engaged sounding responsive at different relative positions in the dialogues. The positions are percentage of all caller speech in the dialogue.*

Dialogue position	Percentage active
0-24	55%
25-49	25%
50-84	13%
85-100	22%

Figure 1 shows how the manually tagged pitch contours occurred at different phases of the dialogue. As can be seen response tokens with rising pitch were most common in the first part of the dialogues.

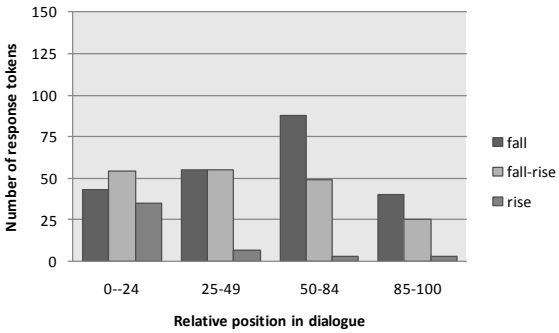


Figure 1. *Distribution of pitch contours over time.*

In order to ensure that the pitch and power measurements on the response tokens were correct, the 31% of all tokens that were produced in overlap with the caller were removed before automatic acoustic analysis was performed. This means that we have automatically analysed approximately 300 response tokens in our study.

3.2. Signal Processing

We use the ESPS pitch tracker and logarithmic power function in the SNACK toolkit with default parameters which gives a 10ms frame rate. From now we refer to log power as intensity. Since the manual labels origin from a single channel with two speakers, special cautions has to be taken at the beginning and end of a speaker change. Small errors in the timing of the labels may cause leakage of a few frames from one speaker to another which makes the pitch tracking inaccurate. Thus, a threshold of 3 consecutive voiced frames is applied at the speaker edges; otherwise the frames are thrown away. In addition, the first and last voiced frame is always omitted. The F0 values are then converted to semitones. Any unvoiced frames between voiced frames are interpolated over using splines. Then a median filter with a 3 frame window is applied, followed by a moving average filter with a 5 frame window. This filtering procedure is applied to both the intensity and to pitch. Each feedback is assigned a parameter *x* which is the elapsed time from the start off the dialog divided by the total dialog duration.

3.3. Clustering of prototypical contours

This study suggests a data-driven intonation model based on a modified length invariant cosine transform (DCT). Each contour f(n) with N points is parameterized by

$$c(k) = \frac{1}{N}\sum_{n=1}^{N} f(n)\cos\left(\frac{\pi}{N}\left(n-\frac{1}{2}\right)(k-1)\right) \quad (1)$$

Both the pitch and intensity contours can effectively be parameterized by a few coefficients with this method. It should be noted that the normalization makes the parameterization independent of speaking rate. This representation was shown to produce better extrapolation than a conventional polynomial parameterization.

We want to find prototypical contours as a function of x, which is the response token's relative position in all caller speech contributions in the dialog. Instead of uniformly split x into beginning, middle and end, an automatic clustering method is proposed. Initially, one feature vector per feedback is constructed by using the first K DCT coefficients for F0 and intensity. The feedback length is also added to the vector. We use K = 3 for monosyllabic and K = 5 for bisyllabic tokens.

Then vector quantization is performed by sweeping x in steps of 0.05, from x_{start} = 0 to x_{end} = min(0.9, x_{start} + 0.75). Let the minimum of the average distortion be found at x_{min}, then the instances $x_{start} < x < x_{min}$ form a cluster and the process is restarted at $x_{start} = x_{min} + 0.1$ with an updated x_{end}. The number of clusters is chosen such that all significant minima in average distortion are found. The look ahead of 0.75 is chosen to avoid splits at high x values for the first clusters. The centroids (mean values) are transformed using inverse DCT, stretched to the average duration (the last feature vector value) and plotted in Figure 2.

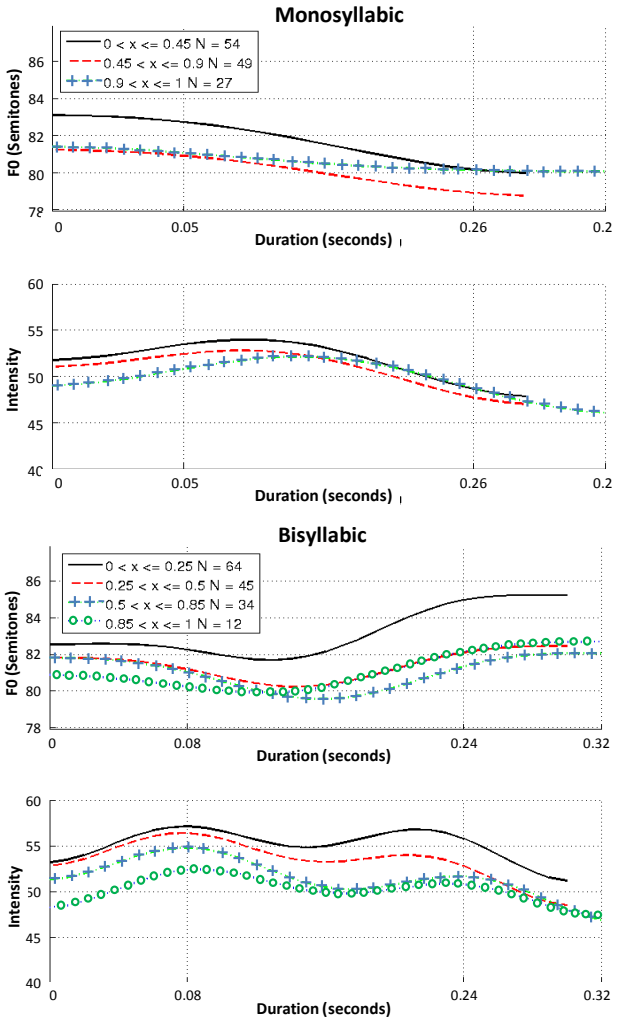


Figure 2: *Pitch and intensity curves as a function of the relative position in dialog. Monosyllabic feedback at the top and bisyllabic at the bottom.*

To further characterize each prototypical curve, the mean pitch and intensity along with pitch slope are calculated for the prototypes and given in Table 3. It should be noted that computing mean pitch and intensity from the actual data points is mathematically equivalent to computing the first DCT coefficient. As can be seen in Figure 2 generally monosyllabic feedback tokens have a falling pitch slope. Over time both the mean intensity and the mean pitch decreases. In the initial phase the second pitch peak is higher in the bisyllabic response tokens. Later in the dialogue both pitch peaks become more even in height, with a smaller pitch movement. The intensity of both peaks fall ever time; where the second intensity peak falls more: The effect of this is that the bisyllabic responsive in the last phase get have more intensity on the first pitch peak which contributes to make the responsive sound less engaging.

Table 3. *Measurements of prosodic prototypes.* x_{start}/x_{end} *are boundaries of relative position in dialog for each prototype.*

Syll.	x_{start}	x_{end}	mean F0	F0 slope	mean Int.
1	0.00	0.45	81.7	-20.0	51.7
1	0.45	0.90	80.1	-5.9	50.7
1	0.9	1.00	80.6	-7.3	49.7
2	0.00	0.25	83.2	11.6	55.1
2	0.25	0.50	81.4	2.8	53.4
2	0.50	0.85	81.0	1.7	51.4
2	0.85	1.00	81.2	8.1	50.1

4. Prosodic alignment

Humans that engage in dialogue have been found to align to their interlocutor's conversational behavior in many respects, e.g. body posture, facial expressions, choice of lexical tokens and prosody [13]. In the current study, prosodic alignment in intensity and pitch slope were investigated. Initial experiments confirmed our assumption that prosodic alignment only occur for some of the callers. We decided to only investigate prosodic alignment of intensity level and pitch slopes on the last syllables. Thus, we pick the last syllable in the caller IPU that preceded a feedback and compare it with the last syllable in the feedback token. Then the intensity level and pitch slope of the last syllable in the feedback is predicted via linear regression from the intensity level or pitch slope of the last syllable in the preceding caller IPU. This was done for all dialogues with four or more non-overlapping non-lexical response tokens. Three points is the minimum number to detect deviation from a straight line and a higher number will cause too much loss of data. By choosing four points enough data was available for doing a more detailed analysis. This left 34 dialogues and 217 response tokens. The dialogues for which alignment was detected were merged into an alignment set each for intensity level and pitch slope. The criteria was quite modest, $R2 > 0.2$, $p < 0.2$ (F-test) and $\beta(2) > 0$. The condition on β encodes the basic assumption of similar behavior rather than opposite behavior. The alignment sets are further divided into start, $x < 0.33$; middle, $0.33 < x < 0.66$ and end; $x > 0.66$. We also examine the effect for mono- and bisyllabic feedbacks.

The intensity alignment set contained 10 of 34 dialogues and 56 of 217 feedbacks, while the pitch alignment set only contained 6 of 34 dialogues and 35 of 217 feedbacks. These numbers indicate that prosodic alignment is a rare phenomenon in these data, especially for pitch. It also raises the question if the result is an artifact of data mining. To test this, a random data set was constructed in which the measured mean intensity or pitch slopes of the non-lexical response tokens and the last part of their preceding turns were replaced by random values. The random data points were sampled from a N(0,1) distribution and the number of dialogues, d_{rand}, for which the prosodic alignment criteria hold was calculated. This process was repeated 10000 times, and the resulting mean and standard deviation for d_{rand} was calculated. Then a right side, one sample, z-test was done for the number of found dialogues in the pitch and intensity alignment sets. The p-value was 0.06 for the pitch set and << 0.01 for the intensity set. Thus, for the pitch set the alignment is near the detectable limit while for the intensity set the found alignment is established. The final results are summarized in Table 4.

Let us first examine the results for the intensity set. The R2 value for all feedbacks is 0.25, so a large part of the variance comes from other factors. For bisyllabic feedbacks the R2 was 0.40 while it was only 0.16 in the monosyllabic case. Thus, the intensity alignment is much stronger for bisyllabic utterances. The intensity alignment also increases

with the relative position in the dialog. It should be noted that since a greater share of the feedbacks at the end of the dialogue are monosyllabic, while most feedbacks at the start are bisyllabic, the overall increase in alignment over dialog can not be explained by the distribution of the number of syllables. This means that alignment actually increases for both monosyllabic and bisyllabic feedbacks over the dialog.

The R2 value for all feedbacks in the pitch set is 0.41, which is higher than for all feedbacks in the intensity set. The R2 of 0.51 for monosyllables is higher than the R2 of 0.26 for bisyllabic feedbacks. This finding is the opposite of what was found in the intensity set. However, one must be careful to jump to conclusions. The pitch contour of bisyllabic feedback is more complex than for monosyllabic ones. Thus, the finding may be explained by the crude parameterization, a simple pitch slope. Let's consider the R2 values for the beginning, middle and end. It starts low at 0.25, then increases in the middle to 0.56 and drops towards the end to 0.37. If the change as a function of position in dialog can be explained by the differences in alignment for mono- and bisyllabic feedback, then there would have been a peak at the end. The drop at the end is contradictorily to the results for the intensity set. However, the drop might have been a peak if bisyllabic alignment would have been more correctly measured. Another possible interpretation is that intensity is aligned more as a habit, regardless of engagement, but pitch slope is more connected to engagement, which explains the drop at the end, but must nevertheless go through a synchronization phase, which explains the low R2 value at the beginning. But the interpretation of the F0 results should have less importance than the intensity results which are more significant.

Table 4. *Prosodic alignment between caller and radio host.*

Intensity alignment set				
Type	**R2**	**F-test**	**N**	**β(2)**
All	0.25	0.00	56	0.3
Mono-syll.	0.16	0.03	29	0.2
Bi-syl.	0.40	0.00	27	0.6
All (start)	0.11	0.13	21	0.2
All (middle)	0.28	0.01	23	0.4
All (end)	0.36	0.04	12	0.3
Intensity non-alignment set				
Type	**R2**	**F-test**	**N**	**β (2)**
All	0.00	1.00	161	0.0
Pitch alignment set				
Type	**R2**	**F-test**	**N**	**β(2)**
All	0.41	0.00	35	0.4
Mono-syll.	0.51	0.00	15	0.4
Bi-syl.	0.26	0.02	20	0.2
All (start)	0.25	0.07	14	0.3
All (middle)	0.56	0.02	9	0.4
All (end)	0.37	0.03	12	0.6
Pitch non-alignment set				
Type	**R2**	**F-test**	**N**	**β (2)**
All	0.00	0.59	182	0.0

5. Conclusions

In this study we have investigated the prosodic patterns of non-lexical response tokens in a Swedish call-in radio show. The professional active listener mostly responded with response tokens at pauses in the callers' speech. In a study of the non-lexical "mm" and "mhm" we have found similar pitch contours as for studies on response tokens in English [8] and Japanese [5]. A rising pitch is associated with interest and encouragement for more speech from the interlocutor, and a feedback with falling pitch functions as acknowledgement and signals lesser interest. For bisyllabic response tokens it is the pitch slope of the loudest syllable that decides which of these two engagement levels the feedback signals. The distribution of feedbacks with different pitch contours changes as a function of dialogue position. The interest-signalling and encouraging pitch contours are most common at the beginning of the call. Over time the mean intensity of the feedbacks decreases, the bisyllabic becomes flatter and the overall pitch level decreases. At the very end this pattern changes where the mean and slope of the pitch increases slightly.

We also tried to find signs of prosodic alignment between the last syllable in the caller's speech and the last syllable in the succeeding response token. For intensity we found signs of alignment of intensity levels in one third of the dialogues, while it was harder to detect dialogues with signs of alignment of pitch slopes. However, the dataset is too small to draw any conclusions, other than that is harder to detect pitch slope alignment.

The implication of our results on conversational speech synthesis is that if we want to synthesize conversational grunts it is not enough to add the sounds of non-lexical response tokens like "mhm" and control the pitch and duration. In order to display the different functions and degrees of interest we also need to be able to control the intensity level continuously on the individual syllables.

6. Acknowledgments

This research is carried out at KTH Speech, Music and Hearing. Funding was provided by the Swedish Research Council (VR) project "*Introducing interactional phenomena in speech synthesis*" (2009-4291).

7. References

[1] Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (pp. 567-578). Chicago.

[2] Fujimoto, D. (2007). Listener Responses in Interaction: A Case for Abandoning the Term, Backchannel. *Journal of Osaka Jogakuin 2year College, 37*, 35-54.

[3] Gardner, R. (2001). *When listeners talk: Response tokens and listener stance.* Amsterdam & Philadelphia: John Benjamins Publishing Company.

[4] Cathcart, N., Carletta, J., & Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In *10th Conference of the European Chapter of the Association for Computational Linguistics.* Budapest.

[5] Ward, N. (1998). The Relationship between Sound and Meaning in Japanese Back-channel Grunts. In *4th Meeting of the (Japanese) Association for Natural Language Processing*.

[6] Pammi, S., & Schröder, M. (2009). Annotating meaning of listener vocalizations for speech synthesis. In *International Conference on Affective Computing & Intelligent Interaction.*. Amsterdam.

[7] Allwood, J. (1995). An activity based approach to pragmatics. In Bunt, H., & Black, B. (Eds.), *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics* (pp. 47-80). Amsterdam: John Benjamins.

[8] Benus, S., Gravano, A., & Hirschberg, J. (2007). The prosody of backchannels in American English. In *Proceedings of ICPhS XVI* (pp. 1065-1068). Saarbrücken, Germany.

[9] Allen, J. F., Ferguson, G., & Stent, A. (2001). An architecture for more realistic conversational systems. In *Proceedings of IUI-0*, 1-8. Santa Fe, NM, US.

[10] Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50*(8-9), 630-645.

[11] Gustafson, J., Heldner, M., & Edlund, J. (2008). Potential benefits of human-like dialogue behaviour in the call routing domain. In *Proceedings of PIT 2008*. Berlin/Heidelberg: Springer.

[12] Couper-Kuhlen, E. (2001). Interactional prosody: High onsets in reasonfor- the-call turns. In *Language in Society* (pp. 29– 53).

[13] Branigan, H., Pickering, M., Pearson, J., & McLean, J. (2010). Linguistic alignment between humans and computers. In *Journal of Pragmatics*.