Cues to perceived functions of acted and spontaneous feedback expressions

D. Neiberg, J. Gustafson

Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Sweden
[neiberg,jocke]@speech.kth.se

Abstract

We present a two step study where the first part aims to determine the phonemic prior bias (conditioned on “ah”, “m-hm”, “m-m”, “n-hn”, “oh”, “okay”, “u-hu”, “yeah” and “yes”) in subjects perception of six feedback functions (*acknowledgment*, *continuer*, *disagreement*, *surprise*, *enthusiasm* and *uncertainty*). The results showed a clear phonemic prior bias for some tokens, e.g “ah” and “oh” is commonly interpreted as surprise but “yeah” and “yes” less so. The second part aims to examine determinants to judged typicality, or graded structure, within the six functions of “okay”. Typicality was correlated to four determinants: prosodic central tendency within the function (CT); phonemic prior bias as an approximation to frequency instantiation (FI), the posterior i.e. CT x FI and judged Ideality (ID), i.e. similarity to ideals associated with the goals served by its function. The results tentatively suggests that acted expressions are more effectively communicated and that the functions of feedback to a greater extent constitute goal-based categories determined by ideals and to a lesser extent a taxonomy determined by CT and FI. However, it is possible to automatically predict typicality with a correlation of $r = 0.52$ via the posterior.

Index Terms: feedback, functions of feedback, goal driven categories, taxonomy

1. Introduction

The nature of the communicative functions of feedback, the proposed categorizations, the cues and the terminology is a subject of considerable debate. The functions of feedback have been proposed to support the grounding process [1] in which participants continuously work at establishing a common ground by signaling perception, understanding, acceptance and a number of emotional states. The opinion on the cues to these functions range from the standpoint of [2] who claims that prosody is decisive, i.e. the tokens are merely carriers of intonation, and [3] who puts more emphasis on instant and incremental phonemic-to-meaning mapping. This leads us to the first part of the current study: To what degree do phonemic realization and prosody contribute to listeners categorization of feedback?

Moreover, if the goal of feedback is to communicate the state of the grounding process and emotional states, it raises the question on how to synthesize these in a dialogue system. For example, one may feed the speech synthesis using cues derived from acted or spontaneous data. While acted stimuli may be perceived more clearly, spontaneous stimuli may be perceived as more natural. We intend to explore this question using methodology from cognitive psychology on categorization [4].

This methodology attempts to access the mechanisms of how humans perceive categories by determine correlates to the typicality of each stimuli. For example, a sparrow is perceived as more typical for birds than an ostrich [5]. Thus, the continuum among categories range from the most typical member to the most atypical members. Three determinants to typicality stands out 1) the members similarity to the central tendency (CT) as measured by co-occurring correlates among members (e.g feathers, color, shape, number of legs, etc.) 2) the members frequency instantiation (FI), i.e. how often they occur as members of their category 3) members similarity to ideals (ID) associated with the goals served by its category.

Categories can be divided into taxonomies (e.g. birds, apples, beers) and goal driven categories (e.g. things to bring on a picnic). The typicality of members in taxonomies are primary determined by their frequency instantiation and secondly by their central tendency and third by ideals. Goal driven categories are primary determined by ideality, secondly by frequency instantiation and third by central tendency. As mentioned, ideals can be thought of as those characteristics that a category member should possess in order to best serve the goals associated with its category. According to the theory of grounding, functions of feedback also serve the goal of communication. This means that the ideal for feedback expressions are to communicate these as efficient as possible. Previous studies has identified emotional facial expressions [6] and vocal expressions [7] as goal driven categories. The latter study also showed that acted expressions are perceived as closer to ones ideals than spontaneous expressions. Thus, acted expressions are communicated more clearly than spontaneous expressions. Finally, if the affective function of feedback is predominant, this would imply that feedback functions are goal driven categories. On one hand, it would not be surprising to find that functions of feedback are emergent processes to support evolutionary shared goal or abstract constructs formed via cultural ideals. On the other hand there is evidence from statistical analysis (which often makes central tendency assumptions) of prosodic cues to feedback functions [8, 9], as well as automatic classification experiments [10, 11], which supports the view that the functions forms a taxonomy.

The current study is divided in two parts 1) the first part examines the interaction between phonemic realization and prosody by determine the phonemic prior bias to different functions 2) the second part is a pilot study which aims to explore determinants to the graded structure within functions.

2. Data

The current dataset was produced in a cooperative project between the speech group at KTH and the speech synthesis company Cereproc[1]. The aim was to record feedback tokens such as “ah”, “m-hm”, “m-m”, “n-hn”, “oh”, “okay”, “u-hu”, “yeah” and “yes” that expressed functions like acknowledgment, continuer, disagreement, surprise, enthusiasm and uncertainty. These functions partly corresponds to the functions derived from a survey where subjects were asked to judge

[1] http://www.cereproc.com/

functional similarity of spontaneous occurring feedback tokens, without giving any directions on which functions to use [9]. In a first recording session all tokens were given a dialogue context to act out, that would help the professional voice-over artist[2] to produce all feedback token with all expressions, like:

GL: Continue for about three blocks and pass the opera.

PH: Yeah. *uncertain*

This gave a combination of 54 types of expressions, each uttered three times, which in total gave 162 vocal expressions.

In a second phase the two were recorded while playing chess, and in a third recording they engaged in socializing conversation. The latter two phases elicited feedback tokens with expressions that partly overlapped those recorded in the scripted session. The feedback tokens were identified and annotated according to their function. Since we wanted to use acoustic measurements for determining the graded structure, we opted to select only one type of token to avoid bias in the measurements from differences in phonemic realization. We selected "Okay" since its' abundant occurrence in the spontaneous expressions and since it had a decent spread among categories and felt rather neutral as a carrier.

Table 1: *Categories of "Okay".*

Value	Acted	Spontaneous	Total
acknowledgment	3	3	6
continuer	3	5	8
disagreement	3	0	3
enthusiastic	3	0	3
surprise	3	1	4
uncertainty	3	3	6
Total	18	12	30

2.1. Prosodic analysis

A common metaphor of studying the communicative aspect of emotions is the Brunswikian lens model [12]. It describes a process which starts with an encoding stage of emotional expression which changes a number of acoustic features of the voice - the distal indicators - for example fundamental frequency (F0). In the receiving party these are decoded as proximal percepts, i.e. the F0 is perceived as pitch, and then an emotional "gestalt" is formed in cerebral cortex.

Our previous study on the encoding stage showed that the functions were expressed with a rather contrastive prosody [13]. *Enthusiasm* and *surprise* showed a higher average F0, as well as shorter duration which also *acknowledgment* showed. The function of *continuer* showed rising F0 which was contrastive to all other functions. The least contrastive functions, *disagreement* and *uncertainty*, only differed in M-F0, while *surprise* and *enthusiasm* differed only in spectral CoG. These results are promising since they indicate that the actor was successful in encoding these functions. This study focuses on the decoding stage.

3. Listeners Decoding Ability

This part aims to examine the interaction between phonemic realization and prosody by determine the phonemic prior bias to different functions.

3.1. Method

10 Subjects of various gender (Females = 4, Males = 6) and age (M = 36.6; SD = 12.1) rated all acted stimuli in a forced choice task by answering the question: *"Which category is this? (acknowledgment / continuer/ disagreement/ enthusiasm / surprise / uncertainty)"*. The stimuli were presented in in randomized order; 15 per page and subjects could change their decisions within each page before submitting the data.

3.2. Result

The results for judging the acted categorized via the forced choice task is shown as a confusion matrix in Table 2. The recall rates per category are found across the diagonal and range between 2-3.8 times the chance level which is 17%. There are two main confusion patterns 1) *surprise* is often detected as *enthusiasm* 2) *uncertainty* is often detected as *disagreement*. The recall rates per token, decomposed into the contribution of the different functions, is shown in Figure 1. The recall rates range from 2.2-3.4 times the chance level. However, all functions can be not be equally well decoded for different kinds of tokens. "Yes" and "yeah" are not likely to be decoded as *surprise*. "ah" and "oh" are over-interpreted as carrying *surprise*, but are not likely to be decoded as *uncertainty* or *enthusiasm*. The token with the most even spread in the contributions from the functions (in terms of entropy and descending order) is "m-m", followed by "u-hu", "okay", "m-hm", "yes", "yeah", "n-hn", "ah" and "oh".

Table 2: *Decoders confusion matrix. The functions are abbreviated as sur: surprise, unc: uncertainty, dis: disagreement, con: continuer, ent: enthusiastic and ack: acknowledgment.*

True	Detected					
	ack	con	dis	ent	sur	unc
ack	**64**	10	7	3	12	4
con	21	**56**	4	1	5	12
dis	22	13	**41**	1	4	20
ent	16	4	1	**52**	26	0
sur	21	6	1	35	**35**	1
unc	19	17	24	0	6	**34**

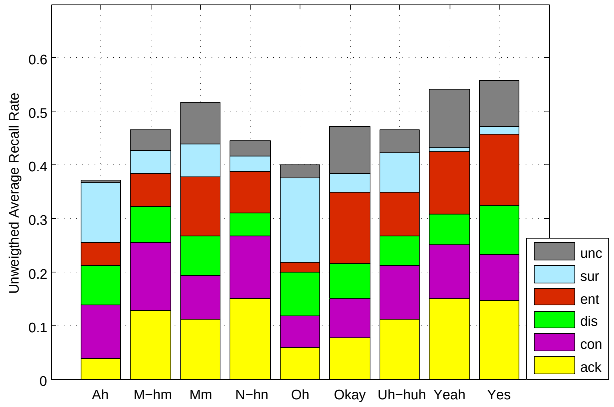

Figure 1: Recall rate per token type. The contribution of the different functions to the recall rate is given within each bar. The functions are abbreviated as sur: surprise, unc: uncertainty, dis: disagreement, con: continuer, ent: enthusiastic and ack: acknowledgment

[2] Paul Hamilton, http://www.pajh.org/acting/index.html

3.3. Discussion

The results indicate that subjects had difficulties in discriminating *surprise* from *enthusiasm*, and to some extent *uncertainty* from *disagreement.* This was expected, since our prosodic analysis on this material showed smaller prosodic differences between the confused functions. The present study complements the previous results by showing the existence of a phonemic prior bias: "ah" and "oh" tend to be strong carriers of *surprise* but not for *enthusiasm* or *uncertainty*. Similarly, "yeah" and "yes" are weak carriers of *surprise*, "mm" is the most *neutral* token and other fall somewhere between. This points towards an interaction between the phonemic surface realization and prosody for the sound-to-meaning mapping. This phonemic prior bias might arise from the subjects experiences of how frequent a certain token is used to express particular function. The phonemic prior bias is related to the frequency instantiation which has been shown to be an important determinant for the typicality within categories (cf. [4]). At this stage we hold the standpoint that the prior is at least one component of FI.

4. Determinants to the Graded Structure

This part is a pilot study which aims to explore determinants to the graded structure to the functions of "okay".

4.1. Method

In the second part, all stimuli were present in random order on a single page. For each stimuli, the subjects were asked the following questions (cf. [7]):

Typicality : *How typical is the expression for [category] in feedback? (1-10)*

Ideality : *If someone want to express [category] in feedback, how effective would this vocalization express [category]? (1-10)*

Condition :*"Is this expression acted or spontaneous? (acted / spontaneous)"*

were *[category]* is the associated function. The ratings were obtained by the same subjects as in the first study.

Instead of obtaining the central tendency from the time consuming process of pairwise judgments, we compute it from prosodic measurements. We use the ESPS pitch tracker and logarithmic power function in the SNACK toolkit with default parameters which gives a 10ms frame rate. The F0 values are converted to semitones and log power is referred to as intensity. Any unvoiced frames between voiced frames are interpolated over using splines. The F0 and intensity trajectories are parameterized using a type II DCT modified by dividing the coefficients with the duration of the token (estimated from the first to the last voiced frame). There are two main reasons for using this time-varying parameterization: 1) The DCT basis functions are periodic which allows good interpolation of syllabic rhythm in speech. 2) The length-invariance gives a normalization for duration or speaking rate. This makes it possible to consider duration separately in the analysis. This parameterization has been used successfully for classification, [14, 15], visualization [16] and has shown to have modest correlation with judged similarity [9]. For this task, a time resolution of 4 coefficients is used. The final feature vector is composed of 4 DCT coefficients of F0, 4 DCT coefficients of intensity, token duration and spectral center of gravity. To obtain appropriate weightings of the dimensions, the pairwise distances is computed in a space rotated via linear discriminant analysis (LDA) were the priors are set to an uniform distribution to avoid correlation to FI. Since the rotated space maximizes the distances between categories, the measurements for central tendency will be sensitive to variance between categories but not to invariance between categories. CT is then computed as the average distance to all other members of the category. As mentioned, using acoustic measurements instead of pairwise judgments has the advantage of being less time consuming and more objective, but on the other hand, there is no guarantee that hidden prosodic variables are present. Obtaining Frequency Instantiation (FI) is not without problems. In previous studies, this determinant was obtained by letting subjects judge FI directly by relying on their experience. As pointed out by [7], such subjective judgments may not reflect the actual frequencies. Instead we use the recall rates for the functions as transmitted by "okay" as determined by the first study.

Bayes theorem can be interpreted as the posterior probability is proportional to the prior of a parameter (e.g. the frequencies of different functions of feedback as transfered by "Okay") multiplied by the likelihood (i.g. a function of CT). Bayes formula is commonly used in statistical classifiers (e.g Naive Bayes, LDA or Hidden Markov Models) and formalizes the relation between FI and CT. By transforming the CT into an approximation to a likelihood, $l(CT|function) = \exp(-(CT)^2)$, one can compute an approximation to the posterior probability. This determinant, the posterior, is important from a affective computing perspective since it gives an indication on how methods used in machine learning can predict typicality.

4.2. Result

The recall rate for correctly judging stimuli as acted was 53% and for spontaneous it was 55%, which is only slightly above the chance level of 50%. The ICC(C,k) [17] (i.e. Cronbach's alpha) was 0.67 for typicality and 0.78 for ideality. The average values were saved for the successive analysis and are shown in Table 3. The ratings for typicality was higher for acted than for spontaneous expressions, and the same for ideality, but there was no significant difference between typicality and ideality for acted functions, and no difference between typicality and ideality for spontaneous functions (*t*-tests, all $p < 0.05$). The Pearson correlations to the determinants of typicality and the cross-correlations between them are shown in Table 4. Due to the sparseness of data, we do not present separate correlations for acted and spontaneous conditions.

Table 3: *Average ratings for typicality and ideality for acted and spontaneous conditions, ranged between 1-10.*

	Typicality	Ideality
Acted	7.04	6.98
Spontaneous	5.78	5.69

4.3. Discussion

The typicality of the functions of "okay" were found to be best predicted by their suitability to express a certain function, i.e. similarity to ideals, and secondly by prosodic central tendency and phonemic prior bias. This indicate that feedback functions to a greater extent are goal driven categories and to a lesser extent form a taxonomy. While subjects could barely determine directly whether an expression was acted or spontaneous, the results was more contrastive when ideality and typicality was judged. This shows that acted expressions are more effectively

Table 4: *Pearson correlations between mean ratings of typicality (TP), ratings of ideality (ID), prosodic measurements of central tendency (CT), the phonemic prior bias and the posterior approximation for prior X likelihood. All correlations are significant to the* $p < 0.02$ *level.*

	TP	ID	CT	Prior	Posterior
TP	**1.00**	**0.96**	**-0.49**	**0.45**	**0.52**
ID		1.00	-0.49	0.50	0.55
CT			1.00	-0.54	-0.77
Prior				1.00	0.95
Posterior					1.00

communicated. The higher ideality of acted expressions suggests that what corresponds to ones ideals is less often found in spontaneous speech. For an analogy to this, consider that the ideality of a low-calorie diet is zero-calories, but zero-calorie food is not very common even in low-calorie diets. The slightly higher correlation for the posterior determinant suggests that using a simple statistical classifier which weight FI and CI according to Bayes theorem is a decent but not perfect choice for applications which attempt to mimic human cognitive processing of the communicative functions. Overall, these results follows [7], although the present study presents a higher CT and a lower FI. The former difference may be due to the objectively measured CT, while to latter difference may be due using phonemic prior bias as an approximation to FI. However, the results of this part of the study must be taken with caution. The number of stimuli is limited and the results are only derived from the token of "okay". Although the proposed method for determine CT is objective, one cannot exclude the presence of hidden prosodic variables. Approximating FI by phonemic prior bias has the advantage of more precisely showing what FI constitutes, however, there might be other components to FI.

5. Conclusions

The present study examines the decoding stage in the Brunswikian lens model [12] of feedback functions and complements our previous study on the same material for the encoding stage [13]. That study showed that the similarity in prosodic realization makes it hard to distinguish *enthusiasm* from *surprise* and *uncertain* from *disagreement*. However, the current study shows that by making use of the phonemic prior bias, confusion could be avoided. When a system is supposed to convey *surprise* it should make use "ah" and "oh" feedback tokens, and when it needs to communicate *enthusiasm.* it should use "yeah" and "yes" tokens. Similarly, "yeah" has a better chance to be recognized as *uncertain*, while "oh" more often gets recognized as *disagreement*. If the system wants the feedback function to be more vague it should make use of the more *neutral* feedback tokens "m-m" and "okay".

When examining the graded structure within the functions of "okay" in the present study it was tentatively suggests that feedback functions to a greater extent are goal driven categories and to a lesser extent form a taxonomy. However, it is still possible to automatically predict typicality with a correlation of $r = 0.52$ via the posterior. Finally, it was found that acted expressions are more effectively communicated. Depending on the situation a dialogue system might need to be more or less clear in its feedback. In some situations it might be sure what it wants to communicate - in these situations it should opt for acted feedback tokes with a strong bias like "oh" and "yes". In more unclear situation the system might want to keep a straight face by producing a feedback token with less clear function - in these situations the system should make use of tokens like 'm-m" and "okay", preferably taken from real interactions rather than acted.

6. Acknowledgements

Funding was provided by the Swedish Research Council (VR) projects 2009-4291 and 2009-4599.

7. References

[1] J. Allwood, J. Nivre, and E. Ahlsen, "On the Semantics and Pragmatics of Linguistic Feedback," *Journal of Semantics*, vol. 9, no. 1, pp. 1–26, 1992.

[2] D. Bolinger, *Intonation and its uses: Melody in grammar and discourse*. London: Arnold, 1989.

[3] N. Ward, "Non-lexical conversational sounds in American English," *Pragmatics and Cognition*, vol. 14, no. 1, pp. 129–182, 2006.

[4] L. W. Barsalou, "Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories," *Journal of experimental psychology Learning memory and cognition*, vol. 11, no. 4, pp. 629–654, 1985.

[5] E. Rosch and C. B. Mervis, "Family resemblances: Studies in the internal structure of categories," *Cognitive Psychology*, vol. 7, no. 4, pp. 573 – 605, 1975.

[6] G. Horstmann, "Facial expressions of emotion: does the prototype represent central tendency, frequency of instantiation, or an ideal?" *Emotion*, vol. 2, no. 3, pp. 297–305, 2002.

[7] P. Laukka, N. Audibert, and V. Auberge, "Exploring the determinants of the graded structure of vocal emotion expressions." *Cognition emotion*, no. August 2011, pp. 37–41, 2011.

[8] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in american english," in *Proceedings of the 16th International Congress of Phonetic Sciences 2007*, 2007, pp. 1065–1068.

[9] D. Neiberg, J. Gustafson, and S. Giampero, "Semi-supervised methods for exploring the acoustics of simple productive feedback in swedish," *Speech Communication*, submitted.

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

[11] D. Neiberg and J. Gustafson, "The prosody of swedish conversational grunts," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, sep 2010, pp. 2562–2565.

[12] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *Eur. J. Soc. Psychol.*, vol. 8, p. 467487, 1978.

[13] D. Neiberg and J. Gustafson, "Towards letting machines humming in the right way - prosodic analysis of six functions of short feedback tokens in english," in *Fonetik 2012*, Göteborg, Sweden, jun 2012.

[14] ——, "Predicting speaker changes and listener responses with and without eye-contact," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, Florence, Italy., sep 2011.

[15] D. Reidsma, I. de Kok, D. Neiberg, S. Pammi, B. van Straalen, K. Truong, and H. van Welbergen, "Continuous interaction with a virtual human," *Journal on Multimodal User Interfaces*, vol. 4, no. 2, pp. 97–118, jul 2011.

[16] J. Gustafson and D. Neiberg, "Prosodic cues to engagement in non-lexical response tokens in swedish," in *DiSS-LPSS Joint Workshop 2010*, Tokyo, Japan, sep 2010.

[17] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients: Correction," *Psychological Methods*, vol. 1, no. 4, p. 390, 1996.