Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 451–469

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Semi-supervised methods for exploring the acoustics of simple productive feedback

Daniel Neiberg *, Giampiero Salvi, Joakim Gustafson

Dept. of Speech, Music and Hearing, KTH Royal Institute of Technology, Lindstedtsv. 24, 100 44 Stockholm, Sweden

Received 16 January 2012; received in revised form 14 December 2012; accepted 14 December 2012
Available online 7 January 2013

Abstract

This paper proposes methods for exploring acoustic correlates to feedback functions. A sub-language of Swedish, *simple productive feedback*, is introduced to facilitate investigations of the functional contributions of base tokens, phonological operations and prosody. The function of feedback is to convey the listeners' attention, understanding and affective states. In order to handle the large number of possible affective states, the current study starts by performing a listening experiment where humans annotated the functional similarity of feedback tokens with different prosodic realizations. By selecting a set of stimuli that had different prosodic distances from a reference token, it was possible to compute a generalised functional distance measure. The resulting generalised functional distance measure showed to be correlated to prosodic distance but the correlations varied as a function of base tokens and phonological operations. In a subsequent listening test, a small representative sample of feedback tokens were rated for *understanding*, *agreement*, *interest*, *surprise* and *certainty*. These ratings were found to explain a significant proportion of the generalised functional distance. By combining the acoustic analysis with an explorative visualisation of the prosody, we have established a map between human perception of similarity between feedback tokens, their measured distance in acoustic space, and the link to the perception of the function of feedback tokens with varying realisations.
© 2013 Elsevier B.V. All rights reserved.

Keywords: Social signal processing; Affective annotation; Feedback modelling; Grounding

1. Introduction

In everyday conversation, listeners usually give brief audio-visual feedback cues while the conversational partner is talking. These tokens have been referred to as signals of continued attention (Fries, 1952), accompaniment signals (Kendon, 1967), verbal listener responses (Dittmann and Llewellyn, 1968) and back-channels (Yngve, 1970; Duncan, 1974). Over time, at least 20 terms with altering definitions have been suggested (Fujimoto, 2007). The vast terminology related to listener feedback calls for providing a careful definition of the tokens of interest in the current study. Listener feedback can be either visual (head nods, eyebrow movements or facial expressions) or verbal. The analysis of verbal feedback by Duncan and Fiske (1977) differentiates between short feedback (words like "okay") and long feedback (phrases like "yeah that's right"). There are also a number of more elusive and primitive feedback tokens like "uhu" and "mm", that are sometimes referred to as non-lexical feedback. However, in the current study we introduce the term *simple productive feedback* for a subset of these tokens. This is inspired by the study on Swedish feedback by Allwood (1987), that describes a set of simple base tokens (mainly sonorants) onto which a set of prosodic and phonological operations are applied to generate verbal feedback tokens with different meanings. This results in a productive combinatorial system that can give rise to a large number of feedback variations. Other studies on the Swedish verbal feedback system include: an early taxonomy of their functions (Sigurd, 1984), perceived effects of synthesized feedback with varying prosody

* Corresponding author. Tel.: +46 (8) 790 7567; fax: +46 (8) 790 7854.
E-mail address: neiberg@speech.kth.se (D. Neiberg).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.12.007

(Wallers, 2006), a comparison between Swedish and Italian feedback (Cerrato, 2006) attitudes in responsive cue phrases (Hjalmarsson, 2010; Neiberg and Gustafson, 2010), prosodic shape as a function of engagement (Gustafson and Neiberg, 2010), issues in annotation and prosodic elicitation (Edlund et al., 2009), usage in dialogue systems (Bell and Gustafson, 2000; Gustafson et al., 2008) and timing (Heldner et al., 2011).

The term feedback indicates a receiving role in the conversation. By continuously giving feedback on what is being said the active listener participates in the collaborative communicative process. Clark's theory of grounding (Clark and Schaefer, 1989) describes discourse as a joint activity in which participants continuously aim to establish a common ground. In this grounding process the listener can relate to what has been said using questions, clarifications and feedback tokens. Allwood et al. (1992) describe four basic communicative functions of feedback that corresponds to Clark's grounding levels (Clark, 1994). These are willingness and ability to (a) continue the interaction (b) perceive the message (c) understand the message (d) accept the message, including attitudinal reactions. Apart from agreement the latter include affective expressions such as interest, surprise, certainty and enthusiasm – all expressed via prosody and salient wording. Theories on emotion include families of basic emotions (Ekman, 1992), epistemic mental states (Baron-Cohen, 2004), expressions described by the component process model of appraisal theory (Scherer, 2009) and the continuous space of valence and arousal in constructivist emotion theory (Russell, 2003). The addition of emotions and mental states makes the list of possible functions of feedback very long. In a study on the perceived functions of listener vocalisations Pammi and Schroder (2009) found that e.g. "aha" could be used to convey several epistemic mental states: *certainty*, *agreement*, *interest* and *anticipation*. This means that claims concerning the contributions of different cues to the functions will be stronger if the analysis is not dependent on a specified set of functions. This is one of the methodological challenges that the current study addresses by performing an initial open class analysis. In the subsequent closed class analysis the following five functional scales were used: *understanding*, *agreement*, *interest*, *certainty* and *surprise*. These were selected since they were considered important functions of feedback when building artificial conversational partners.

The function and meaning of conversational contributions are usually annotated with Dialogue Acts. The proposed ISO standard for Dialogue Acts[1] lists nine dimensions of communicative function for a discourse event. Eight of these are primarily social in nature, concerning dialogue management issues and interpersonal feedback, and they rely more on rich prosodic information than on text content. Manual annotation of dialogue Acts involve judgment of the intention of the speaker, which is a non-trivial task. A prerequisite of assessing the meaning of feedback tokens is to detect them. A characteristic trait of feedback is their short durations, which means that duration alone often can be used to discriminate them from other types of spoken contributions. (Edlund et al., 2010b; Neiberg and Truong, 2011). Even though feedback contains a limited set of words and non-verbal vocalisations, it can be hard to distinguish them from fillers or other types of responses. In order to avoid the problem of defining feedback, Edlund et al. (2009) proposed the operational concept Very Short Utterances (VSU), which is defined as speech segments shorter than 1 s which do not contain high-content elements (nouns, verbs and adjectives) or extra linguistic sounds. *The intra-speaker context criteria for feedback tokens* are that they are short and contain a limited set of sonorants and words.

Yngve (1970) noticed that feedback is common in overlapped speech. He put forward the idea of a main half-duplex channel in conversation, which meant that overlapped speech including feedback has to be transmitted in a back-channel. The frequent occurrence of feedback in overlapped speech has been observed in studies on turn-taking (Sacks et al., 1974; Schegloff, 2000) which led to definitions of feedback as utterances which are not full turns (Ward and Tsukahara, 2000). Empirically feedback has been found to be over-represented in overlapped speech for English (Çetinand Shriberg, 2006; Neiberg and Truong, 2011) and Swedish (Neiberg and Gustafson, 2011a). *The cross-speaker context criteria for feedback tokens* is that they are over-represented in overlapped speech and they are often preceded by prosodic, visual and syntactic cues from the interlocutor (Duncan, 1972; Goodwin, 1981; Ward and Tsukahara, 2000; Gravano et al., 2012).

The productive form of feedback tokens and the definition of the subset used in the current paper are presented in Section 1.1, and Section 1.2 outlines the methodological pathway of investigating the mapping between their form and function.

1.1. Simple productive feedback

We aim to investigate verbal feedback – tokens that are usually not part of syntactical constructions and that are only found in conversation. According to the phonological and morphological analysis by Allwood (1987), feedback in Swedish can be systematised as *primary feedback* like "ja" and "okej", *secondary feedback* like "oj" (oh) and "fy" (ouch) and *simple base morphemes* like "a" and "m" and other sonorants. There are unique phonological operations on feedback that are used on the sonorant part to create bi-syllabic versions. The reduplication is achieved with:

prosodic marking – mm, jaa, nää
insertion of h – mhm, jaha, nähä
insertion of glottal stop – m'm, ja'a, nä'ä

[1] ISO DIS 24617-2 (2010) Language resource management – Semantic annotation framework (SemAF), Part 2: Dialogue acts.

The syllabification is more distinct for glottal stop and insertion of "h", than for prosodic marking where tokens like "mm" and "jaa" are realised with mid-segment intensity drops or double pitch peaks. According to Allwood there are also phonological operations that are not unique for feedback tokens: reduction of consonants and vowel addition. Feedback is also found to undergo general reduplication operations (Moravcsik, 1978), like syllabic reduplication for emphasis ("jaja"). Common Swedish morphological operations that are also applied to feedback include compoundification ("m-okej") and derivation ("oj-san").

According to Ward (2006) non-lexical feedback tokens can be said to form a sub-language of American English that has the purpose of regulating the interaction in conversation, as well as expressing attitudes and emotions. In the current study, we define a feedback sub-language of Swedish, **simple productive feedback**, as: *short feedback where the base tokens are observed to regularly undergo all the reduplication operations that are unique for feedback.* This leads to a set of simple feedback tokens that can be realized in many variations to convey attention and attitude in dialogue. A similar system of feedback has also been put forward by Stromqvist and Richthoff (1999) who documented feedback usage among infants and hypothesised that these play an important role when infants learn to speak. A neurocognitive hypothesis is that the sonorants that make up simple productive feedback are processed by specialised parts of the brain, which allows for a more frequent interjection into overlapped speech (Neiberg and Gustafson, 2012b).

In the current study, we merely wanted to investigate to what degree reduplication operation, base token and prosodic realization give rise to the perception of feedback functions. The result of these investigations will give insight into the importance of the reduplication operations in relation to the base tokens they are applied to. It will also make it possible to investigate how different prosodic realisations of the simple productive feedback tokens influence their perceived function. We chose to define "simple productive feedback" in order to avoid using "infected" terms like "feedback morphemes" or "non-lexical feedback".

1.2. Methodological pathway

There are two goals with the current study. Firstly, it aims to describe context-independent phonological and prosodic correlates to functions of commonly occurring feedback token in Swedish conversations, following conventional methodologies used in studies on vocal affect. Secondly, it proposes methods to achieve the first goal while maintaining a high degree of automation to ensure both speed and objectivity.

All procedures are based on three working hypotheses on the cues for transmitting feedback functions:

Hypothesis 1. Phonological operations change the functions independent of the base tokens.

Hypothesis 2. Base tokens have inherent functions.

Hypothesis 3. Prosodic cues (pitch, intensity and duration) change the functions independent of the feedback tokens.

It should be pointed out that the independence assumption is merely methodological and based on the assumption that interlocutors do not spontaneously produce two cues which stand in conflict. The hypotheses will be evaluated in the following three functional perception tests:

Test 1: Does a change from "m" to "mhm" change the function in the same way as "a" to "aha"?
Test 2: Does a change from "m" to "a" change the function in the same way as "mhm" to "aha"?
Test 3: Does a change in pitch pattern change the function for all feedback tokens in a consistent way?

Dietrich et al. (2006) found that both prosodic and phonemic cues are important for affective recognitions rates for interjections with high lexical content, while recognition rates for the low-lexical category depends more on the prosodic rendering. Due to the low-lexical nature of simple productive feedback, we intend to examine hypothesis no 3 more carefully.

The large number of possible feedback functions leads to a methodological challenge. In order to handle this, a listening experiment was designed where humans annotated the conveyed similarity of feedback tokens with different prosodic realisations and base tokens. Instead of asking for specific functions they were asked to judge conveyed similarity to a reference token. By selecting a set of stimuli that had varying prosodic distances from the reference token, it was then possible to compute a generalised functional distance measure. By not forcing named functions onto the different stimuli, the preconceptions of the subjects on how the functions "should" sound were avoided, and the subjects could focus on the salience of the prosodic realisations and used base tokens. In order to capture the strategies the subject used in these comparisons they were asked to fill in a survey, where they were asked what they listened for when executing the task. By compiling the most common functions given in the survey and accounting for the functions postulated by theory, a closed class analysis could then be performed.

The entire procedure is shown schematically in Fig. 1. There are three steps towards the final goal, each leading to a reduction in data:

Semi-automatic annotation – A subset rich of short utterances was selected from of a corpus of human-human conversations. By using a decoder which objectively models phonemic/prosodic form and cross-speaker dependencies, a semi-automatic annotation could be performed (see Section 3).
Open class analysis – A combination of automatic and human-driven clustering was performed to get a set of

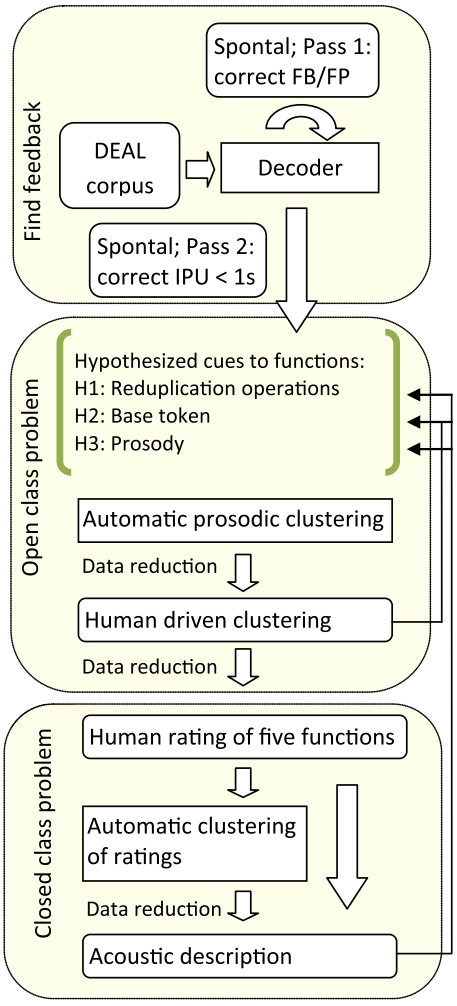

Fig. 1. Methodological pathway. The hypotheses are used in data reduction and verified in listening tests. FB = feedback, FP = filled pause, and IPU = inter pausal unit.

OPEN CLASS CENTROIDS. These were the basis in a listening test that investigated how base tokens and prosodic realisation influence the perceived feedback functions (Hypotheses 2 and 3, see Section 4).

Closed class analysis – The OPEN CLASS CENTROIDS derived from the clustering were rated for: *understanding*, *agreement*, *interest*, *surprise* and *certainty* (see Section 5). Correlation, regression and ANOVA analysis was performed to investigate the contribution of base tokens, phonological operations and prosodic marking (Hypotheses 1–3). The OPEN CLASS CENTROIDS and were further clustered according to the functional ratings resulting in a smaller set of CLOSED CLASS CENTROIDS. This created a hierarchy described from top to bottom as consisting of closed class clusters whose corresponding closed class members were also OPEN CLASS CENTROIDS with their own open class members. The relationship between form and function in this hierarchy was examined using an exploratory visualisation technique.

2. The dialogue corpora

2.1. The DEAL corpus

The initial detector for *filled pauses* and *feedback* used in the current study (Neiberg and Gustafson, 2011a), was trained on the DEAL corpus of eight task-oriented role-playing dialogues (Hjalmarsson, 2008). The DEAL corpus was chosen because it is had been annotated for filled pauses and feedback, and had been found to be rich in such items. The domain was conversational second language training in a flea market scenario. The face-to-face dialogue corpus used 6 subjects (4 male and 2 female), 2 posing as shop keepers and 4 as potential buyers. The two dialogue partners were given conflicting tasks: the customers were assumed to try to lower the price of each item under discussion, while the shop keepers were instructed to try to get as much as possible.

2.2. The Spontal corpus

The main experiments presented in the current paper were conducted on the Spontal corpus (Edlund et al., 2010a), which consists of recordings of spontaneous face-to-face socialising conversations in which the participants were given no directions regarding task or topic of conversation. The corpus contains 120 half-hour dialogue sessions, roughly divided into three 10 min sections, indicated to the subject by a brief comment over the intercom from the recording leader. The subjects were not given any instructions on what to talk about. However, after approximately 20 min they were asked to pick up and open a wooden box that was placed on the floor between them. They could then choose to discuss the content of the box or continue the on-going topics of conversation. Only the first two blocks of each dialogue were used in the current study to get a more coherent style of interaction. The subjects were all native speakers of Swedish and balanced for gender, the interlocutor's gender and as to whether they knew each other. In order to select a subset of dialogues that were most probable to be rich in feedback tokens, all dialogues were analysed by the voice activity detector described in (Heldner et al., 2011), and the resulting

Table 1
The dialogues which had the highest proportion of IPUs shorter than 1 s were selected. S# = session number, F. = friends, and G. = gender.

S#	F.	G. 1	G. 2	IPU < 1 s (%)	Minutes
06	Y	F	M	54	24.0
29	N	M	M	54	25.1
02	N	F	F	54	25.8
26	N	F	M	57	15.9
10	N	F	M	58	24.0
17	Y	M	M	62	25.0
05	Y	F	F	63	13.7
32	N	M	F	66	22.4

speech/silence patterns were used to rank the dialogues according to the proportion of inter pausal units (IPUs) shorter than 1 s. The frequencies ranged between 26% and 66%, suggesting that the feedback strategies were highly individual. The eight dialogues with highest proportion of short IPUs were selected. Some properties of the dialogues and interlocutors are shown in Table 1.

3. Semi-automatic annotation

Semi-automatic transcription is expected to reduce labour work considerably for large corpora and rely on more objective criteria than using only human judgments. In order to achieve this, a decoder has to be trained to model the essential characteristics of the target dialogue acts. In (Neiberg and Gustafson, 2011a), a stochastic decoder for *feedback*, *filled pauses*, *silence* and *speech* was proposed and optimised on the DEAL corpus. The decoder was based on Coupled Hidden Markov Models with Gaussian mixtures as emitting distributions capable of modelling the following characteristics of feedback:

Feedback is usually isolated – Transitions such as *silence-feedback-silence* or *silence-feedback-speech* were probabilistically coded via the first order Markov assumption.
Feedback is short – Durational modelling was encoded via self-transition probabilities.
Feedback consists of sonorants and specific words – Phonemic and acoustic realisations were captured via MFCCs and a prosodic cepstrum representation at 50 ms frame rate.
Feedback is overrepresented in overlapped speech – A joint transition matrix for a dual Markov chain, one per speaker, captured the cross speaker dependencies.

A common problem in conversational corpora is cross-talk. This was suppressed by using the Joint Maximum Cross-Correlation (JMXC) feature, i.e. the maximum cross-correlation between channels (Laskowski et al., 2004). To avoid over-fitting and keep the recall rate high, we used a configuration with a lower number of model parameters (referred to as CDT) than the one that is optimal for the DEAL corpus.

3.1. Annotation procedure

In order to get an initial automatic detection of *filled pauses* and *feedback* in the selected Spontal dialogues, the decoder trained on the DEAL corpus was used. To speed up the process, only the detected *filled pauses* and *feedback* tokens were manually corrected by two experts in a forced choice task. This meant that *filled pauses* and *feedback* token that had been detected as *speech* or *silence* were not corrected. Extra-linguistic sounds such as lip- and tongue-smacks and hawks were consistently annotated as *extraling*, and the tokens "eh", "em" and "hm" as *filled pauses*.

The inter-label agreement was 0.72 (accounting for the marginal distribution and a Cohens kappa of 0.64, $p < 0.01$). The hit rates for the decoder given the agreed labels are shown in Table 2. This gives the quality of the detected *filled pauses* and *feedback* tokens. The result shows that the recall rate is acceptable but the precision rate is poor. This may indicate a mismatch in dialogue style and recording conditions between the DEAL and Spontal corpora.

In a second step, the Gaussian emitting distributions were adapted using a single Expectation–Maximisation iteration. This adaptation process used both the uncorrected *speech/silence* items, and the manually corrected items that had been agreed to be labelled as *filled pauses/feedback/speech*. The retrained decoder was used to again detect segments of *speech*, *filled pauses* and *feedback* in the selected Spontal dialogues. In a subsequent manual inspection all detected *filled pauses*, *feedback* and *speech* segments shorter than 1 s were manually corrected in a forced choice task, where the detected classes were preselected. Since a large portion was already correctly detected, this method decreased the manual labour of labelling the data for *speech*, *filled pauses* and *feedback* significantly. The resulting inter-labeller agreement was 0.80 (accounting for the marginal distribution and a Cohens kappa of 0.72, $p < 0.01$) and the hit rates for the decoder, given the agreed labels, are shown in Table 3. The evaluation criterion became harder in the second pass, due to the addition of the short *speech* segments to detected feedback and filled pauses. Despite this, the precision increased dramatically, while the recall rate remained high.

Finally, the agreed labels were manually annotated according to their phonemic content; see Table 4, where doubled letters like "aa" indicates bi-syllabic tokens. Most of the *feedback* tokens defined by Allwood (1987) were found. This included simple productive feedback variations as well as lexical feedback such as "okej" (okay) and "visst" (sure).

Since the goal was to extract context-independent acoustic correlates to feedback functions, we opted to select a

Table 2
Decoder performance after the first pass, given the agreed human annotation of detected feedback and filled pauses.

	Count	Recall	Precision	*F*-score
Feedback	1023	0.95	0.26	0.41
Filled pause	39	0.64	0.03	0.05

Table 3
Decoder performance after the second pass, given the agreed human annotation of feedback and filled pauses and speech segments shorter than 1.0 s.

	Count	Recall	Precision	*F*-score
Feedback	1675	0.95	0.64	0.77
Filled pause	122	0.77	0.28	0.41
Speech	2890	0.63	0.96	0.76

Table 4
Token count for the final annotation.

N	Token
451	m
389	(j)a
80	nä
70	jaha
56	okej
55	mm
47	(j)aa
46	mhm
40	jo
26	nej
24	näe
19	aha
16	ha
11	oj
11	nähä
66	Other productive variations
7	Other single words
16	Extraling
47	Silence
198	Multiple words

subset of tokens that are common in dialogue, and that occur with different functions that are not determined by the previous context. The most common feedback tokens were monosyllabic, where the base tokens "(j)a" and "m" were almost equally common. These base tokens were also equally common in their bisyllabic versions "(j)aa" and "mm". This led to the decision to limit the further analyses to *simple productive feedback* tokens which were variations based on "m" and "a". In the Swedish feedback "ja", the initial "j" is typically reduced to different degrees, often until it is not pronounced at all. This indicates that it bears little meaning in feedback tokens, which is why "ja" and "a" were grouped into "(j)a" and "jaa" and "aa" into "(j)aa" in the current study. However, in order to test this hypothesis "jaha" and "aha" were kept in separate categories.

Since the current corpus contained very few bi-syllabic *feedback* tokens resulting from reduplication with insertion of glottal stops, these were omitted in the current study. Furthermore, in the current study segmental features, like voice quality or allophonic version (e.g. front or back "a"), were not annotated in the current study. From now on the following abbreviations are used: 1S denotes the monosyllabic versions, 2S the bi-syllabic version obtained by reduplication with prosodic marking, and 2SH the bisyllabic version obtained by reduplication with insertion of "h". Similarly, M* and A* refer to all variations created from the feedback base tokens "m" and "a" respectively. The token counts for the selected tokens are summarised in Table 5.

Table 5
Counts for tokens selected for further analysis.

Base token	Phonological operation		
	1s	2s	2sh
M*	451	55	46
A*	389	47	89

4. Obtaining the open class centroids

Manual annotation of the communicative functions of feedback is a hard and tedious task, due to the large number of functions they are used for: turn regulation, grounding and display of the listener's affective and attitudinal state. Furthermore, there is no fixed set of functions, they can co-occur and they can be conveyed to different degrees. A feedback token can for example give back the turn while conveying that the listeners agrees somewhat, but is uncertain. This results in a large number of possible functions, which essentially makes the annotation an open class problem.

Given that prosody is a strong cue to the communicative functions, selecting suitable candidates among all the different realizations as found in data is a challenge. A shortcut to this problem is to extract a few tokens, the OPEN CLASS CENTROIDS which compactly represent functions that have distinctly different prosodic patterns. A typical approach to extracting these would be to conduct vector quantization on multivariate prosodic measurements where the partitions are given by the Lloyd's algorithm (Lloyd, 1982). The centroids in the resulting Voronoi tessellation would then cover the variation in the original data set in a condensed form. Even though automatic prosodic measurements and clustering procedures minimises the annotation effort, it cannot be ruled out that any hidden prosodic variables are vital correlates to the conveyed function. Furthermore, there might be a mismatch between the F0 trajectory and the actual perception of F0. As an example, F0 movements in synthesised sequences of "amama" are more salient for the vowels than for sonorants (House, 1990). This can be handled by letting humans assess the functional similarity between tokens with different prosodic realisations and then transform the result into pairwise distances for clustering. The procedure of judging pairwise similarity to obtain central tendency has been used by Barsalou (1985) as a possible determinant of the most typical instance within a category. Although results suggest that central tendency is not the strongest predictor for the typicality of vocal emotional expressions (Laukka et al., 2011), a recent study in which central tendency was computed from prosodic measurements, indicate that what is perceived as typical for a category should be found in the vicinity of the central tendency (Neiberg and Gustafson, 2012a). The present study also used central tendency to obtain the categories in the first place.

To let subjects judge pairwise similarity for N tokens would require $N*(N-1)/2$ comparisons which makes the task unrealistic. This can be avoided by using an approximation to full pairwise clustering. In the current study, this was achieved in a two-step approach: Firstly,

each token found in Table 5 was quantized in a prosodic space giving 15 prosodic centroids. This was assumed to be enough to represent a significant part of the most commonly conveyed affective classes and cognitive states in socializing dialog. Secondly, the prosodic centroids were used as stimuli in a listening experiment where the subject had to judge functional similarity of feedback tokens with different prosodic realisations. Instead of making a full pairwise comparison, each prosodic centroid was compared to four other stimuli pre-selected based on prosodic distance. A sparse pairwise distance matrix was then obtained by averaging the binary decisions from multiple judges. The listening experiment was divided into two parts. In the first part, the WITHIN CATEGORY, only stimuli with the same base token and same phonological operation as the reference were used. In the second part, the BETWEEN CATEGORY, the reference stimulus had a different base token compared to the target stimuli, while the phonological operation was the same (comparing "m" with "a", "mm" with "aa" and "mhm" with "aha"). This design of the listening experiment allows for testing Hypothesis 2 and 3 (stated in Section 1.2), .e. to what extent prosody and base token are cues to the functions of feedback. In addition, the subjects were surveyed to investigate what they listen for while making their judgments. Finally, the prosodic centroids were merged via a sparse approximation of full agglomerative hierarchical clustering to create categories represented by OPEN CLASS CENTROIDS.

Section 4.1 describes the automatic prosodic clustering, Section 4.2 describes the generation of stimuli, the approximation to agglomerative hierarchical, execution of the listening test and the subsequent analysis, and Section 4.3 discusses the results.

4.1. Obtaining the prosodic centroids

4.1.1. Step 1: signal processing

Pitch and intensity were measured using the ESPS pitch tracker and logarithmic power function in the SNACK toolkit with default parameters which gives a 10 ms frame rate. From now on logarithmic power is referred to as intensity. Only tokens with more than 100 ms of voiced frames were kept. The F0 values were then converted to semitones. Any unvoiced frames between voiced frames were interpolated over using splines. Then a median filter with a 3 frame window was applied, followed by a moving average filter with a 5 frame window. This filtering procedure was applied to both the intensity and the pitch.

4.1.2. Step 2: prosodic distance measure

The F0 and intensity trajectories were parameterised using a type II DCT modified by dividing the coefficients with the duration of the token. This made the coefficients invariant to segment length and made it possible to consider duration separately in the analysis. This parameterisation has been used successfully in related tasks (Neiberg and Gustafson, 2011b; Neiberg and Truong, 2011; Reidsma et al., 2011). A resolution of 6 coefficients has been found to be adequate for parameterisation of bisyllabic tokens (Gustafson and Neiberg, 2010). The 0'th coefficient is equal to the arithmetic average and was omitted to avoid speaker dependent bias. The final feature vector was composed of F0 (coefficients 1–5), intensity (coefficients 1–5) and token duration (computed for the connected voiced frames in the segment). The dimensions belonging to one of the three feature types were then z-normalised using a transform where the scale factors were computed for the coefficient of lowest order (that holds most of the variance). Finally, the dimensions belonging to each feature were multiplied with the following heuristic weights: 2 for F0, 1 for intensity and 4 for duration. This ensured the distance measure to produce more intuitive results.

4.1.3. Step 3: clustering

In the clustering phase, the tokens were clustered separately for each type of phonological operation (1S/2S/2SH) using a codebook obtained from Lloyd's algorithm and the Euclidean norm of the prosodic distance measure. The closest token to the center was saved as a prosodic centroid reference.

4.2. Human-driven agglomerative hierarchical clustering

In this method, human judgments are used to define distances for agglomerative hierarchical clustering. This requires a pairwise distance matrix per token type, each requiring $15 * (15 - 1)/2 = 105$ unique distances. However, having humans judge $105 * 7 = 735$ stimuli is not realistic. As an approximation each of the prosodic centroid references was instead compared to 4 selected target instances. The latter were selected to ensure enough variation as determined by the prosodic distance measure.

The algorithm for generating stimuli was: (1) For each reference instance i, compute the distance to all instances j and sort ascendant. (2) Pick slot numbers 2, 4, 8 and 14 in the sorted list as the selected stimulus (the 1st is referring to itself so it is omitted). The doubling of distance allows for a higher resolution between stimuli with similar prosody where the threshold for functional similarity is hypothesised to be found. The instances furthest away (the 15th) were not included since they often were outliers.

This generated sparse distance matrices as approximations to the full distance matrices. The target stimuli were selected in two sets. In the first set, the WITHIN CATEGORY set, the reference centroid feedback token shared both base token and phonological operation with the selected stimuli. In the second set, the BETWEEN CATEGORY set, the selected stimuli still shared phonological operation with the reference, but it had been applied to the other base token (comparing "m" with "a", "mm" with "aa" and "mhm" with "aha"). By merging the two parts, three sparse distance matrices were obtained, one for each of the phonological operations (1S/2S/2SH).

4.2.1. Approximate agglomerative hierarchical clustering

The quality of hierarchical clustering was determined by the Cophenetic correlation coefficient (Sokal and Rohlf, 1962), which measures how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points. Thus, a dendrogram generated from one of the three full pairwise distance matrices can be compared with a dendrograms generated from a sparse distance matrix. This is shown in Table 6 for dendrograms generated using the prosodic distance measure with unweighted average distance (UPGMA) for linkage. For sparse matrices, the UPGMA was computed by ignoring non-existent distances, which prevented these to be linked. It should be noted that the Cophenetic correlation was computed on the full distance matrix, regardless if the dendrogram was generated from a sparse or full distance matrix. It is clear that the dendrograms generated from sparse distance matrices describes the data as good as the dendrograms generated from full distance matrices in terms of prosodic feature distances. So dendrograms generated using perceptual distances over sparse matrices should describe the full data set as well.

4.2.2. Execution of the listening test

The stimuli selection process presented in Section 4.2 generated in total $15 * 7 * 5 = 525$ stimuli. A rough estimate from pilot tests indicated that it would take at least 90 min to perform a listening test. Since the current study would be using untrained annotators in one session this was not a feasible test size. In order to make the perceptual test manageable, the stimuli set was divided into three parts. The 30 subjects were thus divided into three groups that judged one part each. The three groups of annotators were mixed regarding age and gender.

A web-based listening test was set up as a multiple binary-choice task. At the top of each page a play button for the reference centroid token was displayed. Below it was written: "Which of these convey the same as the reference?", followed by four play buttons for the target stimuli. The target stimuli, which were selected to be at different prosodic distances from the prosodic centroid reference, were presented in random order. The subjects could listen multiple times to both the reference and the stimuli, in order to determine which of the stimuli they thought conveyed the same as the reference at the top. The subjects were allowed to select one, several or none of the stimuli.

4.2.3. Results from participants survey

After the experiment the participants filled out a survey. The survey included a question on what they were listening for when deciding if a stimulus conveyed the same as the reference centroid. Out of 30 participants, 8 mentioned various prosodic cues, 8 mentioned that they listened for attitude, emotion or the semantic/pragmatic meaning in general (without exemplifying). 9 gave examples of functions they had identified, e.g. interest, surprise, certainty, uncertainty, agreement, acceptance, confirmation, boredom. 5 mentioned that they imagined a context for the feedback – they first tried to think of a question or dialogue context that could have preceded the reference token and they then imagined that context while listening to each of the 4 stimuli.

Table 6
The Cophenetic correlation for dendrograms generated using sparse or full distance matrices.

Matrix	1S	2S	2SH
Full	0.85	0.92	0.65
Sparse	0.85	0.91	0.68

4.2.4. Explaining judged distances by prosody

The binary decisions from the human judges were converted to numerals indicating distance according to: 0 = same function and 1 = different function. Similarity between the prosodic distance measure and human judges was verified for the binary decisions using t-tests. The null-hypothesis was: there is shorter prosodic distance between the reference and the stimuli that were not conveying the same as the reference. For the WITHIN CATEGORY set, the null hypothesis could be rejected for all judges while for the BETWEEN CATEGORY set the null hypothesis could be rejected for 28 of 30 judges ($p < 0.05$ for left side t-tests).

The averaged binary decisions were first examined as a function of slot number of the sorted prosodic distances. This is shown for the between/within categories sets in Fig. 2. It is clear that the averaged functional distances are proportional to the prosodic distance measure. The averaged distances were further examined by computing correlation coefficients. The overall Pearson correlation was 0.42, while the Spearman correlation was 0.46 (both significant $p \ll 0.01$). The higher correlation for the Spearman coefficient indicates the presence of non-linear terms. The Spearman correlation for the WITHIN CATEGORY was

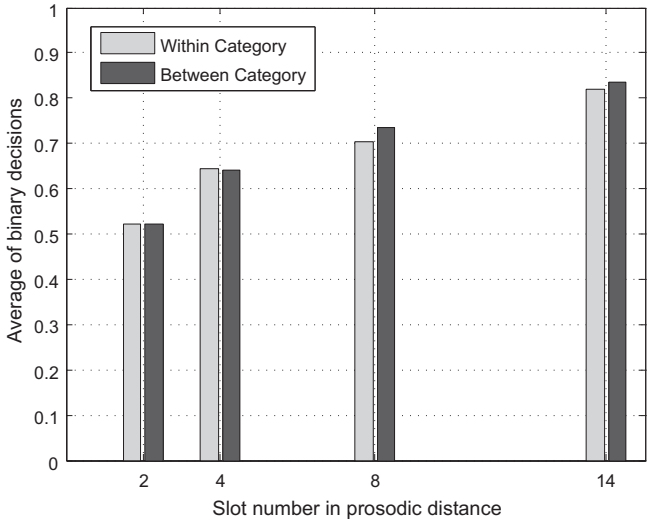

Fig. 2. The average of the binary decisions as a function of slot position in the ascending sorted prosodic distances.

0.46 and for the BETWEEN CATEGORY 0.48, indicating little or no difference between these two sets. For the merged matrices formed from the two sets, the Spearman correlation was 0.56 for 1S, 0.24 for 2S and 0.49 for 2SH, indicating an exception for the 2S type tokens. Further analysis showed that for the 2S tokens in the WITHIN CATEGORY, the low correlation was due to "aa", while similar low correlation was found for the 2S tokens in the BETWEEN CATEGORY. The correlations are summarised in Table 7–9.

4.2.5. Clustering for the open class centroids

The method of approximate agglomerative hierarchical clustering described in Section 4.2.1 was used together with the generalised functional distance to obtain the OPEN CLASS CENTROIDS. It is not straightforward to choose the number of open class clusters that represents a meaningful partition of the data. The criteria are chosen to minimise the number of open class clusters while maintaining prosodic similarity within the clusters. The latter task was performed by a human judge (the first author). For this purpose, a graphical user interface was implemented where the partition of clusters was visualised in a tree structure and where it was possible to listen to the feedback tokens located in the leaf nodes. The cut-off criteria for leaf nodes were based on distance measure rather than the commonly used inconsistency coefficient. This made the adjustments easier, since it corresponds to a straight line through the dendrogram. The human judge would adjust the global cutoff threshold that the system used to generate the cluster tree. The prosodic similarity was then assessed by listening to the feedback tokens in each cluster. If they did not sound similar enough, the global threshold could be adjusted to generate new clusters until the result was satisfactory. The cutoff distances were chosen as $d_{1s} = 0.55$, $d_{2s} = 0.44$, $d_{2sh} = 0.61$. This gave in total 32 clusters out of which 13 contained only one data point. The latter were considered unusual outliers and where removed from any further processing. This gave 6 clusters for 1S, 6 for 2S and 7 for 2SH, in total 19 open class clusters. In each cluster, the data point nearest to the centre was chosen as the OPEN CLASS CENTROID.

4.2.6. Open class purity in respect to base token

In order to investigate to what degree the base tokens contributed to the communicative function, a cluster purity measure was computed with the following algorithm: Each cluster i and member j was assigned a value $s_{ij} = 1$ for A* and $s_{ij} = -1$ for M*. Then the cluster purity with respect to base token is $a_i = 1/N_i|\sum_{1<j<N_i} s_{ij}|$ where N_i is equal to the number of members in each cluster. The value a_i is zero for equal proportion of A*/M* within cluster i and equal to one for only A* or only M* tokens. The average a^i per phonological operation is given in Table 10.

4.3. Implications for the initial hypotheses

The correlations between the human-derived functional distances and the automatic prosodic distance measure clearly show that prosody is a strong factor for the conveyed functions of feedback. However, the low correlation for the "aa" tokens in the WITHIN CATEGORY and when compared with "mm" in the 2S tokens in the BETWEEN CATEGORY indicates that other factors than prosody were dominant cues to the function of "aa". These factors may include allophone variations (front or back "a"), nasalisation or difference in voice quality. Little difference was found in the overall correlation between prosody and function for the WITHIN CATEGORY (same base token) and the BETWEEN CATEGORY (different base token). This is interpreted as support for the hypothesis that prosody changes the function independently of base token (Hypothesis 3 in Section 1.2). Again, other cues than prosody seem important for the function of "aa", which makes the result for 2S less conclusive.

The cluster purity with respect to base token is increasing with phonological complexity, from 1S, 2S to 2SH. This increase cannot be an artefact of the chosen distance thresholds since the number of clusters is equal for the 1S

Table 7
Spearman correlation between derived human distances and the prosodic distance measure for the merged sparse matrices.

Overall	1S	2S	2SH
0.46	0.56	0.24	0.49

Table 8
Spearman correlation between derived human distances and the prosodic distance measure for the merged sparse matrices in the WITHIN CATEGORY set.

Overall	0.46
m	0.47
mm	0.43
mhm	0.57
(j)a	0.59
(j)aa	0.18
aha	0.41
jaha	0.37

Table 9
Spearman correlation between derived human distances and the prosodic distance measure for the merged sparse matrices in the BETWEEN CATEGORY set.

Overall	1S	2S	2SH
0.48	0.58	0.22	0.54

Table 10
Cluster purity with respect to base token.

Operation	1S	2S	2SH
Purity	0.34	0.40	0.54
Clusters	6	6	7
Total no. points	23	26	41

and 2S sets, and almost the same for the 2SH set where the latter actually has more members per cluster. The observation that the salience of the base token increases with phonological complexity gives weak support for the hypothesis of inherent functions of these (Hypothesis 2 in Section 1.2), since there is an interaction with the phonological operation. Also, small difference in the overall correlation between prosodic distance and generalised functional distance for the WITHIN CATEGORY and the BETWEEN CATEGORY indicates that base token is a weak cue.

5. Obtaining the closed class centroids

The hybrid automatic/human-driven clustering procedure resulted in 19 open class clusters. The OPEN CLASS CENTROIDS were selected as stimuli for further analysis since they were considered to be representative for the common functions of the feedback tokens in their clusters. The annotation scheme for these functions is inspired by the Mumin annotation scheme (Allwood et al., 2007). According to this scheme, the relevant functions for feedback are continuation/contact and perception (CP); continuation/contact, perception and understanding (CPU); agreement/acceptance and affect/attitude in general. The functions are proposed to have polar dimensions (cf. Sigurd, 1984) and follow a hierarchy: agreement/acceptance and affect/attitude implies CPU; and CPU implies CP. In the current study, the selected feedback tokens were assessed to be positive along the CP dimension (due to their phonemic realisations). This led to an annotation scheme that consisted of *understanding*, *agreement* and three affect/attitude functions. The selection of these functions was based on the survey from the first listening experiment, resulting in three polar dimensions: *interest*, *surprise* and *certainty*. These also correlate with the Baron-Cohen's epistemic mental states found for "aha" in the study by Pammi and Schroder (2009),

5.1. Affective/attitudinal rating

Each stimulus was rated on 5 point Likert-scale $[-2,-1,0,1,2]$ along the following functional dimensions:

1. *non-understanding – understanding*
2. *disagreement – agreement*
3. *uninterest – interest*
4. *expectation – surprise*
5. *uncertainty – certainty*

The default setting in the listening test was 0 on all dimensions, which corresponds to a *neutral* function, i.e. only contact and perception. A total of 20 subjects (19–66 years old, 7F/13M) rated each stimuli according to the five dimensions. In order to facilitate comparison and reconsideration, all stimuli were presented on a single web page.

The ICC(C, k) (McGraw and Wong, 1996) (i.e. Cronbach's alpha) were between 0.91 and 0.96 for the five dimensions and the average values were saved for the successive analysis. A Principal Component Analysis shows that 95% of the variance can be explained by three dimensions, as illustrated in a Pareto plot shown in Fig. 3 and the first two components are shown in Fig. 4. The correlations between the rated dimensions are shown in Table 11. The correlations showed that *understanding* and *agreement* have a strong correlation with *certainty*, while *understanding* and *agreement* only have medium correlation to each other. *Interest* has weak correlation to both *agreement* and *surprise*. Finally, *surprise* has weak negative correlation to *certainty*.

5.2. Open class distance to closed class distance

One basic assumption is the correspondence between the generalised distance of conveyed function and the ratings of the five functions. To verify this, one distance matrix

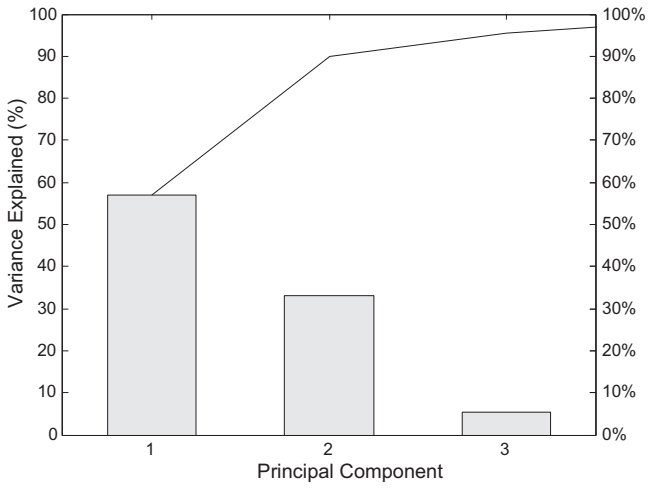

Fig. 3. Explained variance as a function of the principal components found in the five dimensional ratings of feedback tokens.

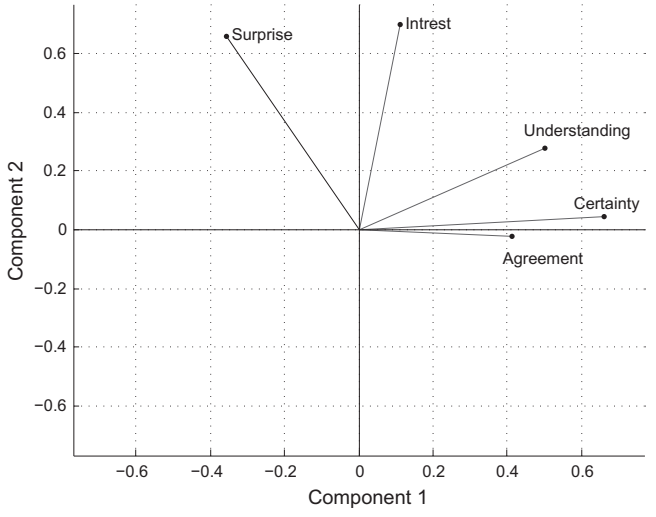

Fig. 4. Loadings of the five rated dimensions of feedback tokens for the first two principal components.

Table 11
Covariance between the dimensions (*: $p < 0.05$, †: $p < 0.01$). Und = understanding, Agr = agreement, Int = interest, Sur = surprise, and Cer = certainty.

	Und	Agr	Int	Sur	Cer
Und	1.00	**0.62**†	**0.48***	−0.22	**0.81**†
Agr		1.00	0.10	−0.43	**0.84**†
Int			1.00	**0.57***	0.22
Sur				1.00	**−0.48***
Cer					1.00

per phonological operation was computed from the functional ratings using Euclidean distance. Each of these matrices can be compared to the distances between the OPEN CLASS CENTROIDS in the sparse distance matrices of generalised function used in Section 4. The two types of distances were compared by cutting out the upper triangles from the matrices and concatenate the sequences for the three phonological operations. The Pearson correlation between the two types of distances is 0.54 ($p = 0.02$). This indicates that the specified functions form a significant proportion of the components of the generalised distance measure. Since the generalised distance was correlated with the prosodic distance measure that was used to select stimuli, we can presume that prosody plays a vital part for the perceived function of feedback.

5.3. Visualisation of the feedback stimuli

In order to explore the relationship between the prosodic form and the perceived function the prosodic realisations of the feedback tokens were visualized. The pairwise Euclidean distances between vector-valued ratings were examined using agglomerative hierarchical clustering and unweighted average distance (UPGMA) for cluster linkage, similarly to approach of Stocksmeier et al. (2007). The distance threshold was set to create seven closed class clusters and their corresponding CLOSED CLASS CENTROIDS as indicated in the dendrogram shown in Fig. 5. It is noteworthy that there is a hierarchy described from top to down as consisting of closed class clusters in which each corresponding closed class members were OPEN CLASS CENTROIDS with their own open class cluster members rated according to the generalised distance measure.

In order to investigate how prosodic patterns found in the 19 stimuli correlate with the closed class ratings and the generalised functional distance, a variation of the exploratory visualisation technique introduced in (Gustafson and Neiberg, 2010) was adopted. In this technique, the line widths of the F0 contours indicate the intensity level, see Fig. 6. In the current study two types of normalisations were applied (1) average of the F0 curves were *z*-normalised per dialogue and speaker, then scaled back using the global standard deviation (2) the intensities, as displayed by line width, were scaled using the global minimum and maximum values for the stimuli. Now recall that each closed class cluster member is a former OPEN CLASS CENTROID, with its corresponding open class cluster members.

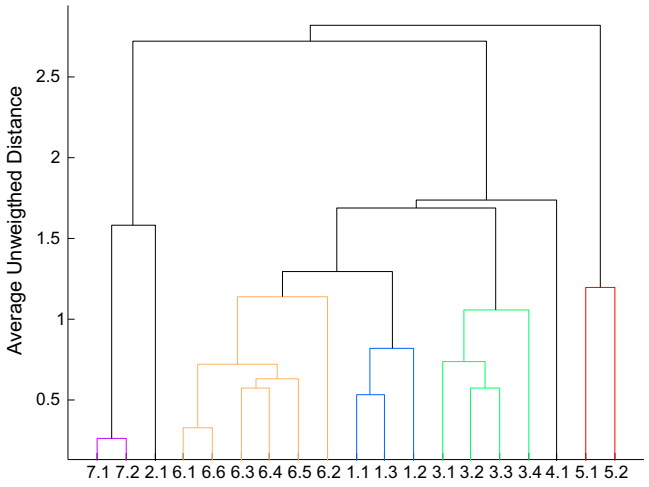

Fig. 5. Hierarchical agglomerative clustering of the OPEN CLASS CENTROIDS using the ratings of the five functional dimensions. The colouring indicate cluster assignment which corresponds to the rows in Fig. 6 where the fundamental frequency contours are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To further investigate the relationship between prosodic form and perceived function, the average contour of each *open class cluster* is also in shown in the figure. This was computed by taking the average of the length invariant DCT coefficients and durations for all members followed by applying inverse transform. An analysis of the prosodic curves and conveyed functions by the groups of feedback tokens (G1–G7) in Fig. 6 lead to the following observations:

- The open class cluster and OPEN CLASS CENTROID share prosody.
- There are tokens that have low ratings on all dimensions which indicate only a neutral (CP) function.
- Feedback tokens are often multi-functional – some tokens convey *understanding*, *agreement*, *certainty*, and negative *surprise* (G1), while other signal *understanding*, *interest* and *surprise* (G5).
- Some functions can be achieved by all tokens – "mhm", "aha", "a" and "m" convey moderate levels of *understanding*, *agreement* and *certainty* (G6).
- Some tokens are connected to certain functions – "(j)aha" conveys *surprise* (G5 and G7) and "m" is often used for grounding on the lowest level (G3).
- As hypothesized, the initial "j" in "jaha" does not lead to any difference in meaning compared to "aha" (G5 and G7).
- Prosodic cues are connected to certain functions:
  - A fast speaking rates and moderate F0 variation lead to moderate ratings for *understanding*, *agreement* and *certainty* (G6).
  - A fast speaking rate and a high F0 with a flat or falling contour lead to high ratings on *understanding*, *agreement* and *certainty* (G1).
  - A fast speaking rate and a moderate F0 rise lead to *understanding* and *interest* (G4).

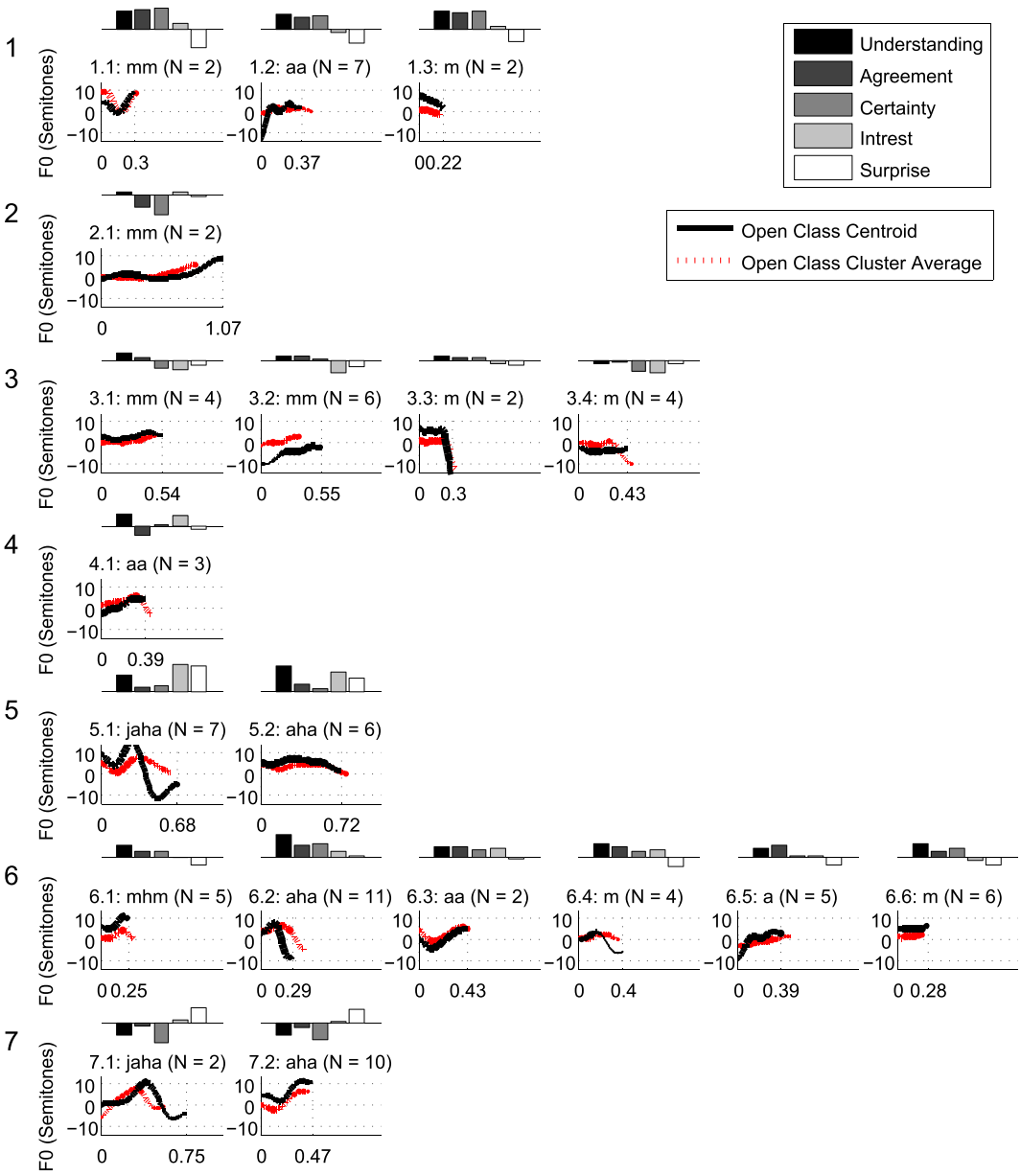

Fig. 6. Ratings of five functions and fundamental frequency contours arranged according to the clustering shown in Fig. 5. The main y-axis shows the number of the obtained CLOSED CLASS CENTROID. The labels "$X.Y$ token N=" above each graph stands for X = CLOSED CLASS CENTROID, Y = closed class cluster member which the same as OPEN CLASS CENTROID, token = phonemic realisation of the OPEN CLASS CENTROID and N = number of open class cluster members. The line widths of the F0 contours indicate the intensity level.

- A moderate speaking rate and medium F0 with a flat contour are perceived as *neutral*, *uninterested* or *uncertain* (G3).[2]
- A moderate speaking rate and a F0 rise on the loudest syllable lead to *uncertainty* and *non-understanding* (G7).
- A very low speaking rate and a slow bisyllabic F0 rise lead to *non-agreement* and *uncertainty* (G2).

In summary, prosody was found to be very important when assessing the functions of simple productive feedback tokens. The fact that feedback tokens often were rated similarly for *understanding*, *agreement* and *certainty* indicates that these are hard to distinguish between. This supports the idea that these functions represent epistemic states to some extent, and that they are strategically used to encourage the other to continue speaking without having to clearly indicate which level the grounding is done at.

5.4. Statistical analysis of acoustic cues to the selected functions

The prosodic cues were examined using a standard set of prosodic measurements, denoted as SET 1 and computed as:

[2] The final drop in 3.3 is due to a pitch tracking error.

duration, per speaker z-normalised average intensity and F0, as well as the slope and standard deviation of F0. Duration was computed for connected voiced segments to avoid any bias from extra silence included in the automatic segmentation. The contribution of each acoustic measurement was determined via acoustic correlates in Section 5.4.1 where the OPEN CLASS CENTROID ratings were reused for each non-centroid member in the respective open class cluster. This bootstrapping is considered safe since the generalised functional distance is correlated with the ratings of the specified functions and since the visualisation procedure shows a similarity of form between the OPEN CLASS CENTROIDS and the open class cluster average. To more formally verify the re-usability of the ratings, the three most salient features where compared for the OPEN CLASS CENTROIDS and their non-centroids as a function of positive and negative affect/attitude ratings in Section 5.4.2. The predictability of the functional dimensions based on acoustic features was evaluated using linear regression analysis in Section 5.4.3. Two feature sets were compared: the standard set, SET 1, and the feature set used for the prosodic distance measure (see Section 4.1.2), here denoted as SET 2. The latter is more suitable for online processing since it is not dependent on speaker dependent z-normalisations and has the advantage to be an orthogonal parameterisation. The contribution of base token and phonological operation was examined using analysis of variance (ANOVA) in Section 5.4.4.

5.4.1. Acoustic correlates to functions

Pearson correlations for the prosodic features of SET 1 are shown in Table 12. The statistically significant correlations show that *certainty* and *agreement* were expressed with similar prosody, while the other functions were expressed with more contrastive prosodic patterns. Furthermore, the signs alone of the significant correlations were found to determine all functions but *certainty* and *agreement*. Finally prosody was found to change the function independently of base token, and prosodic patterns were salient cues to all functions except for *certainty* and *agreement* which both displayed the same cues.

5.4.2. Acoustic features of centroid and non-centroid members in each open class cluster

To verify the feasibility of reusing the ratings for each member in the respective open class cluster, the three most salient features; duration, average F0 and the slope of F0, were examined for positive and negative affect/attitude ratings for the centroids and the non-centroids members of each open class cluster. The results for the functions which were significantly correlated to the three cues are shown in Fig. 7. For all the significant correlations in Table 12, both the centroids and non-centroids average values show the same sign of difference between positively and negatively rated clusters, e.g. positive ratings for *surprise* show higher average F0 for both centroids and non-centroids compared to negative rating for *surprise*.

Table 12
Pearson correlation for each acoustic feature and rated dimensions (*: $p < 0.05$, †: $p < 0.01$). Dur = duration, M-F0 = mean F0, Δ-F0 = F0 slope, and SD-F0: standard deviation of F0.

	Dur	M-F0	Δ-F0	SD-F0
Understanding	0.08	**0.25***	**−0.35**†	−0.07
Agreement	**−0.23***	0.10	**−0.24**†	−0.05
Interest	**0.47**†	**0.41**†	−0.06	0.20
Surprise	**0.50**†	**0.29**†	0.09	**0.25***
Certainty	**−0.29**†	0.17	**−0.31**†	−0.06

5.4.3. Regression analysis

The predictability of the functional dimensions from acoustic features was evaluated using a linear regression analysis with interaction terms. The two prosodic sets, SET 1 and SET 2, were compared using the adjusted R^2 measure for goodness of fit as shown in Table 13. It was shown that SET 2 gave a better fit for all functions except for *interest* for which the two sets produce nearly equal results. The better result for SET 2 is likely due to its ability to model syllabic structure. If only SET 2 is considered, all functions are predicted well in the range $0.48 < R^2_{adj} < 0.67$ except *agreement* which showed poorer fit $R^2_{adj} = 0.35$ with lower significance.

5.4.4. ANOVA of base tokens and phonological operations

An analysis of variance of the ratings for the base token and re-duplication factors are showed in Table 14 on the last page of this article. The statistically significant effects ($p < 0.05$) for both *interest* and *surprise* were found to be the effect of phonological operation and the interaction effect (reduplication X base token).

The results for Tukey–Kramer post hoc tests are summarised in Table 15. The ratings for both *interest* and *surprise* were higher for the 2SH type tokens than for the 2S type tokens and for the 1S type tokens.

5.5. Implications for the initial hypotheses

The results can first be directly related to the initial Hypotheses 1 and 3 (see Section 1.2): In agreement with Hypothesis 1 it was found that 2SH changed the function independently of base token and was a cue for *surprise* and *interest*. In agreement with Hypothesis 3 it was found that prosodic cues changed the function independently of phonemic realisation. Specifically, *understanding* was correlated with increasing average F0 and decreasing slope of F0; *certainty* was correlated with shorter duration and decreasing slope of F0 and the same for *agreement* but with weaker significance; *surprise* and *interest* correlated with longer duration and higher average F0 and *surprise* had the additional cue of increasing standard deviation of F0. In addition, the regression analysis gave that the ratings for all functions could be predicted from prosodic cues, expect for *agreement*. Hypothesis 2, the contribution of base token, was not supported with any statistically

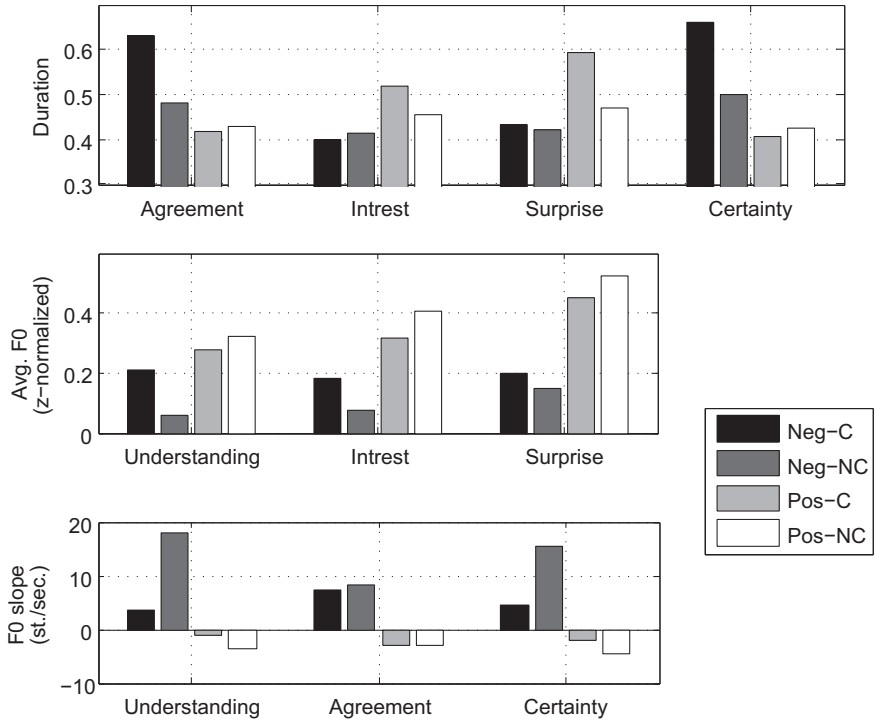

Fig. 7. The three most salient prosodic features for ratings quantized into positive and negative ratings for OPEN CLASS CENTROIDS (C) and non-centroids (NC).

Table 13
Adjusted R^2 measures for linear regression using interaction terms. SET 1 is a classic parameterisation and SET 2 is length invariant DCT coefficients. (*: $p < 0.05$, †: $p < 0.01$).

Function	SET 1	SET 2
Understanding	0.21†	0.67†
Agreement	0.12	0.35
Interest	0.49*	0.48*
Surprise	0.44†	0.63†
Certainty	0.21†	0.51†

significance. However, Hypothesis 2 was weakly supported by the observation that "m"/"mm" tend to be neutral, i.e grounding on the lowest level (CP).

5.6. Acoustic correlates to perceived function

There are few previous studies that have described the acoustic correlates to the function of Swedish feedback tokens. In order to be able to compare the current results with previous research this section cites studies on prosodic cues to the investigated functions across languages and in other types of speech.

5.6.1. Surprise

Surprise can be decoded across cultures in faces (Ekman, 1972) and in vocal expression (Levitt, 1964; Scherer et al., 1991; Sauter et al., 2010a). In general, *surprise* is cross-culturally expressed via high F0 and large F0 variation while the consensus for speaking rate and intensity is weak (Pell et al., 2009; Sauter et al., 2010b; Chen et al., 2004). Thus, the encoding and decoding of *surprise* seems to exhibit some universal acoustic correlates across cultures. Expression of *surprise* has been studied for affirmative cue words (lexical feedback) in American English by (Lai, 2009). Via Likert-scale correlates, *surprise* was found be expressed by high average F0, large F0 standard deviation, longer duration and high average intensity. Except for intensity, this is in agreement with the findings for feedback reported here, i.e. long duration, high average F0 and large F0 standard deviation. In a recent study on the perception of acted and spontaneous emotions on English feedback with 9 different base tokens, it was found to be hard for the subjects to distinguish between *surprise* and *enthusiasm* (Neiberg and Gustafson, 2012a). Both shared prosodic realisation (high average F0 and short duration), and differed only in spectral center of gravity. However, a phonemic prior bias was found: "ah" and "oh" were found to be strong carriers of *surprise*, but weak carriers of *enthusiasm* or *uncertainty*, while "yeah" and "yes" were found to be strong carriers of *enthusiasm* or *uncertainty*, but weak carriers of *surprise*. The effect of *surprise* on vocalisation in terms of a novelty appraisal check is discussed in (Scherer, 1986). It is argued that the inspiration centre is stimulated in order to facilitate information processing of the stimuli which interrupts on-going vocalisation by a sudden inhalation, followed by prolonged exhalation sounds in case of positive surprise. The post hoc analysis in the current study showed that the most characteristic type of tokens for

Table 14
ANOVA analysis of base tokens, phonological operations and their interaction.

Source	S.S.	df.	MS	*F*-val.	*P*-val.
Understanding					
Phonological operation	0.09	2	0.04	0.08	0.93
Error	49.78	87	0.57		
Base token	0.89	1	0.89	1.60	0.21
Error	48.97	88	0.56		
Interaction	4.52	5	0.90	1.67	0.15
Error	45.34	84	0.54		
Agreement					
Phonological operation	0.33	2	0.16	0.72	0.49
Error	19.81	87	0.23		
Base token	0.06	1	0.06	0.28	0.60
Error	20.07	88	0.23		
Interaction	1.08	5	0.22	0.95	0.45
Error	19.05	84	0.23		
Interest					
Phonological operation	14.29	2	7.14	20.20	0.00
Error	30.77	87	0.35		
Base token	0.47	1	0.47	0.94	0.34
Error	44.58	88	0.51		
Interaction	15.01	5	3.00	8.40	0.00
Error	30.04	84	0.36		
Surprise					
Phonological operation	31.00	2	15.50	57.79	0.00
Error	23.34	87	0.27		
Base token	2.06	1	2.06	3.46	0.07
Error	52.28	88	0.60		
Interaction	31.35	5	6.27	22.90	0.00
Error	22.99	84	0.27		
Certainty					
Phonological operation	1.07	2	0.53	1.00	0.37
Error	46.42	87	0.53		
Base token	0.25	1	0.25	0.47]	0.50
Error	47.23	88	0.54		
Interaction	3.47	5	0.69	1.33	0.26
Error	44.01	84	0.52		

Table 15
Tukey–Kramer post hoc statistics for the Likert-ratings $[-2,-1,0,1,2]$ of interest and surprise. M = mean and SD = standard deviation. Significant differences are reported for factors using 95% confidence intervals.

Factor	*M*	SD	Diff. to factor
Interest			
2SG	0.65	0.09	1S, 2S
2S	−0.18	0.12	2SG
1S	−0.11	0.12	2SG
mhm	0.80	0.17	mm, aa, m, a
aha	0.58	0.11	mm, aa, m, a
mm	−0.28	0.17	mhm, aha
aa	−0.07	0.17	mhm, aha
m	−0.14	0.17	mhm, aha
a	−0.08	0.18	mhm, aha
Surprise			
2sg	0.65	0.08	1S, 2S
2s	−0.55	0.10	2SG
1s	−0.50	0.11	2SG
mhm	0.52	0.15	mm, aa, m, a
aha	0.71	0.08	mm, aa, m, a
mm	−0.53	0.15	mhm, aha
aa	−0.56	0.15	mhm, aha
m	−0.51	0.15	mhm, aha
a	−0.49	0.16	mhm, aha

expressing *surprise* was found to be of type 2SH, which also is a voiceless glottal fricative that interrupts the on-going vocalisation.

5.6.2. Interest

The vocal expression of *interest* is not well-studied. Acoustic correlates and measurements have not shown to be very contrastive (Goudbeek and Scherer, 2010; Banse and Scherer, 1996). However, successful regression to Likert-scales for *interest* from multiple acoustic features is reported in (Banse and Scherer, 1996) and cross-cultural automatic classification of *interest* has shown to yield recall rates over chance level (Neiberg et al., 2011). In an analysis of the prosodic cues to *interest/engagement* in non-verbal feedback tokens of a Swedish radio host in a call-in-show, *engagement* was found to be correlated with high intensity and rising pitch (Gustafson and Neiberg, 2010). At the end of each call the radio host's engagement level decreased, which he indicated to the caller by decreasing the overall intensity and F0 standard deviation in his feedback tokens. Liscombe et al. (2003) reported numerous acoustic correlates for interest in English phrases, where significant correlates were: high average F0, large F0 range, high average intensity and shorter syllable length. These results partly agree with the results for feedback reported here: long duration and high average F0. The finding that *interest* is associated with tokens of type 2SH coincides with the finding for *surprise*. By assuming a compositional element of novelty appraisal check in *interest*, this might be explained in the same way as for *surprise*.

5.6.3. Certainty, agreement and understanding

A downstepped pitch has been found to correlate with *certainty* in English (Gravano et al., 2008), and in Swedish clarification ellipses (Edlund et al., 2005). The latter also found that positive *understanding* and *acceptance* were associated with an early peak in F0 followed by a falling pitch slope. This is in agreement with the current study where *certainty* was found to be expressed with falling slope, as well as with shorter duration. *Agreement* was found to share prosodic correlates with *certainty* and consequently their dimensional ratings were highly correlated in the current study. However, when using a regression model for prosody, *agreement* was more poorly predicted than *certainty*. This either indicates a bad choice of regression model, or that segmental cues and voice quality are important when displaying different degrees of (non-)agreement in feedback tokens.

F0 rises have been linked to *uncertainty* in English (Nilsenova, 2006; Reese, 2007), and according to Lai (2010) cue words with a rising pitch indicate that the current question under discussion is unresolved, and that the

listener wants to hear more. Final rises have also been empirically linked to *continuers* in Swedish (Wallers, 2006) and "mm" with final rise has been interpreted as a *continuer* in English (Gardner, 2001). A study on the prosody of *backchannels* showed that they are prosodically marked with higher pitch, intensity and pitch slope than both agreements and other functions (Benus et al., 2007). Stolcke et al. (2000) found that the feedback tokens "right" and "yeah" used as *backchannels* were shorter and had less intensity than those used for *agreement*. The current study cannot verify the results for the prosodic correlates to *backchannels*, since it was not used as a category in the listening test. However, *non-understanding*, *disagreement* and *uncertainty* were found to be correlated with a rising pitch, and *disagreement* and *uncertainty* with a longer duration.

The PCA-analysis of the ratings shows that 95% of the variance can be explained by three dimensions. These fewer uncorrelated dimensions might be interpreted as antecedent appraisal objectives or psychosocial component functions. The projection plot for the first two principal axes and cross-correlation analysis suggest that *agreement* and *certainty* and *understanding* indeed have a common perceived function: to indicate trouble in the communication (Batliner et al., 2003) or to indicate to what degree the current question under discussion is unresolved. The current results are in line with Krahmer et al. (1999) that found prosodic features for "go back" in dialogue to be high boundary tone, long duration and high pitch range.

6. Conclusions

The current paper presented a semi-supervised method for investigating the acoustic correlates to the perceived functions of feedback tokens. Semi-supervised annotation and prosodic clustering were initially used to extract a manageable, but representative set of feedback tokens from a large corpus of human-human conversations. These were then used in two perceptual listening tests, the first aimed to determine the coupling between difference in prosodic realisation and difference in perceived function, and the second test aimed to determine to which degree five selected functions were conveyed by acoustically different feedback tokens.

In order to get a distinguishable set of feedback token, *simple productive feedback* was defined as tokens which purely consist of sonorant and vowels, and that undergo the reduplication operations that are unique for feedback. The selection of these simple tokens made it possible to perform controlled investigations on how the reduplication operations interact with the base tokens and prosodic realisations to form new functions. To verify the usefulness of the proposed method, three hypotheses were introduced (see Section 1.2); The results relation to the initial hypotheses can be summarized as:

Hypothesis 1 : the phonological operation 2SH *do* change the perceived meaning independently of base token, by adding a cue for *surprise* and *interest*;

Hypothesis 2 : base token exhibit *some* inherent functions – overall, the "m"/"mm" tokens are perceived as neutral. Furthermore, allophone variant and voice quality seemed to be more dominant cues to the function of "aa" tokens than of "mm" tokens and base token became slightly more salient as the phonological complexity of the token increases;

Hypothesis 3 : prosodic cues (pitch, intensity and duration) *do* change the function independently of phonemic realisation – prosody was found to be a contrastive cue for all functions, and it was mostly a stronger cue than the base token. However, prosody was not discriminative between *certainty* and *agreement*.

The proposed semi-supervised method has been shown to be able to couple acoustic realisation of feedback tokens to their perceived functions. The correlation between the generalised distance of similarity and the functional ratings was 0.54 ($p = 0.02$), indicating that the specified functions form a significant proportion of the components of the generalised distance measure. By combining the acoustic analysis with an exploratory visualisation of the prosody the cues to the perceived functions could be identified as summarised in Table 16.

The signs of the statistically significant correlations form a prosodic code which discriminate each function, except *certainty* and *agreement*. There is also a complex interaction of phonemic content and prosodic cues. The observed acoustic cues for *surprise* and *certainty* showed agreement with the literature on vocal affect. The shared cue to *non-understanding*, *disagreement* and *uncertainty* is a high positive slope of F0, which is interpreted as a request for the speaker to say something more that might resolve the issue at hand. Predicting the functions from prosodic cues using linear regression analysis showed that the feature set based on length-invariant DCT coefficients (SET 2), was superior in terms of adjusted R^2 compared to a more classic set consisting of average F0/intensity/duration plus F0 slope and standard deviation (SET 1). The goodness of fit was good for all functions ($0.48 < R^2_{adj} < 0.67$ for SET 2) except for

Table 16
Summary of findings. +/− Indicate the sign of acoustic correlates. Dur = duration, M-F0 = mean F0, Δ-F0 = FO slope, and SD-F0 = standard deviation of F0.

	Prosody				Token
	Dur	M-F0	Δ-F0	SD-F0	
Understand		+	−		
Agreement	−		−		
Interest	+	+			2SH
Surprise	+	+		+	2SH
Certainty	−		−		
Neutral					"m"/"mm"

agreement. This indicates that without hearing the context, it is hard to imagine this function of simple productive feedback tokens.

By combining the acoustic analysis with an exploratory visualisation of the prosody, the current paper have established a map between human perception of distance between tokens, the distance between acoustic features, and the link to the perception of attitude. Among many observations, this identified "m"/"mm" as having a neutral function, indicating only contact and perception.

For the computational community, the results presented in the current study may lead to more human-like dialogue systems (Edlund et al., 2008; Gustafson et al., 2008), that achieve instant response via continuous interaction (Kopp et al., 2006; Reidsma et al., 2011; Buschmeier and Kopp, 2011), and that are able to achieve rapport (Gratch et al., 2007; Cassell et al., 2007), social resonance (Kopp, 2010) and conversational grounding (Traum, 1994; Larsson, 2002; Skantze, 2007; Bunt et al., 2007). We believe that achieving grounding in dialogue systems via simple productive feedback is an overlooked opportunity, but yet to be proven in its full potential. In this process it is necessary to be able to generate and understand feedback that indicates the level of *understanding* and *agreement*. This could include rephrasing when rises and longer duration is detected and move on when short duration and negative slope is detected. Detection of feedback that signals non-understanding, uncertainty, disagreement or surprise could be useful to detect misrecognitions (Hirschberg et al., 1999), dis-confirmations (Krahmer et al., 1999) and other forms of trouble in communications (Batliner et al., 2003). This might be handled by initiating error-handling sub-dialogues, by restarting the on-going grounding process or by re-evaluating the current belief state. Detection of *interest* might be useful for assisted browsing systems (Gustafson et al., 2002), artificial companions (Sloman, 2010) and social robots (Payr, 2011). Finally, the findings may be also be useful in tutoring on proper usage of feedback for second language learners (Ward et al., 2007).

This paper is an initial step towards obtaining context-independent phonological and acoustic cues to attitudes of vocal feedback. The next step would be to investigate how the interpretation changes with discourse context. It would be interesting to investigate how the perceived function of a token with a certain prosodic realisation changes as it is placed in different discourse contexts. Personality traits and speaking styles also influence the prosodic realisation of communicative functions. This means that those need to be accounted for when assessing the meaning of feedback tokens – for some people by default produce listening feedback with rising pitch, without intending to convey surprise or lack of understanding. However, this could be handled by storing information about default prosodic patterns in a user model.

Acknowledgments

Funding was provided by the Swedish Research Council (VR) projects "Introducing interactional phenomena in speech synthesis" (2009-4291) and "Biologically inspired statistical methods for flexible automatic speech understanding" (2009-4599).

References

Allwood, J., 1987. Om det svenska systemet för spraklig återkoppling. In: Linell, P., Adelswärd, V., Nilsson, T., Pettersson, P. (Eds.), Svenskans Beskrivning, . In: Tema Kommunikation, vol. 1. University of Linköping, pp. 89–106.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P., 2007. The mumin coding scheme for the annotation of feedback turn management and sequencing phenomena. Language Resources and Evaluation 41, 273–287.

Allwood, J., Nivre, J., Ahlsen, E., 1992. On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9, 1–26.

Banse, R., Scherer, K., 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, 614–636.

Baron-Cohen, S., 2004. Mind Reading: The Interactive Guide to Emotions. Jessica Kingsley Publishers.

Barsalou, L.W., 1985. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. Journal of Experimental Psychology Learning Memory and Cognition 11, 629–654.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nth, E., 2003. How to find trouble in communication. Speech Communication 40, 117–143.

Bell, L., Gustafson, J., 2000. Positive and negative user feedback in a spoken dialogue corpus. In: Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China.

Benus, S., Gravano, A., Hirschberg, J., 2007. The prosody of backchannels in american english. In: Proceedings of the 16th International Congress of Phonetic Sciences 2007, pp. 1065–1068.

Bunt, H., Morante, R., Keizer, S., 2007. An empirically based computational model of grounding in dialogue. In: Proceedings of SigDial, Antwerp, Belgium, pp. 283–290.

Buschmeier, H., Kopp, S., 2011. Towards conversational agents that attend to and adapt to communicative user feedback. In: Proceedings of the 11th International Conference on Intelligent Virtual Agents, Reykjavik, Iceland, pp. 169–182.

Cassell, J., Gill, A., Tepper, P., 2007. Coordination in conversation and rapport. In: Proceedings of the Workshop on Embodied Language Processing, Association for Computational Linguistics, Prague, Czech Republic, pp. 41–50.

Çetin, Özgür, Shriberg, E., 2006. Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition. In: Proceedings of the ICSLP, Pittsburgh, pp. 293–296.

Cerrato, L., 2006. Investigating Communicative Feedback Phenomena across Languages and Modalities. Ph.D. thesis, Dept. of Speech, Music and Hearing, KTH Royal Institute of Technology. Lindstedtsv. 24, 100 44 Stockholm, Sweden.

Chen, A., Gussenhoven, C., Rietveld, T., 2004. Language-specificity in the perception of paralinguistic intonational meaning. Language and Speech 47, 311–349.

Clark, H.H., 1994. Managing problems in speaking. Speech Communication 15, 243–250.

Clark, H.H., Schaefer, E., 1989. Contributing to discourse. Cognitive Science: A Multidisciplinary Journal 13, 259–294.

Dietrich, S., Ackermann, H., Szameitat, D.P., Alter, K., 2006. Psychoacoustic studies on the processing of vocal interjections: how to disentangle lexical and prosodic information? In: Anders, S., Ende, G., Junghofer, M., Kissler, J., Wildgruber, D. (Eds.), Understanding

Emotions, . In: Progress in Brain Research, vol. 156. Elsevier, pp. 295–302.
Dittmann, A.T., Llewellyn, L.G., 1968. Relationship between vocalizations and head nods as listener responses. Journal of Personality and Social Psychology 9, 79–84.
Duncan, S., Fiske, D., 1977. Face-to-Face Interaction: Research, Methods and Theory. Lawrence Erlbaum Associates, Hillsdale, New Jersey, US.
Duncan Jr., S., 1972. Some signals and rules for taking speaking turns in conversation. Journal of Personality and Social Psychology, 23.
Duncan Jr., S., 1974. On the structure of speaker–auditor interaction during speaking turns. Language in Society, 161–180.
Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., House, D., 2010a. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, pp. 2992–2995.
Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A., 2008. Towards human-like spoken dialogue systems. Speech Communication 50, 630–645.
Edlund, J., Heldner, M., Al Moubayed, S., Gravano, A., Hirschberg, J., 2010b. Very short utterances in conversation. In: Proceedings of Fonetik, pp. 11–16.
Edlund, J., Heldner, M., Pelcé, A., 2009. Prosodic features of very short utterances in dialogue. In: Vainio, M., Aulanko, R., Aaltonen, O. (Eds.), Nordic Prosody – Proceedings of the Xth Conference, Peter Lang, Frankfurt am Main, pp. 57–68.
Edlund, J., House, D., Skantze, G., 2005. The effects of prosodic features on the interpretation of clarification ellipses. In: Proceedings of Interspeech 2005, Lisbon, Portugal, pp. 2389–2392.
Ekman, P., 1972. Universals and cultural differences in facial expressions of emotions. In: Cole, J. (Ed.), Nebraska Symposium on Motivation. University of Nebraska Press, Lincoln, Nebraska, pp. 207–283.
Ekman, P., 1992. An argument for basic emotions. Cognition & Emotion 6, 169–200.
Fries, C.C., 1952. The Structure of English: An Introduction to the Construction of English Sentences. Harcourt, New York.
Fujimoto, D.T., 2007. Listener responses in interaction: a case for abandoning the term backchannel. Journal of Osaka Jogakuin 2 year College 37, 35–54.
Gardner, R., 2001. When Listeners Talk: Response Tokens and Listener Stance. John Benjamins Publishing Company, Amsterdam and Philadelphia.
Goodwin, C., 1981. Conversational Organization: Interaction Between Speakers and Hearers. Academic Press.
Goudbeek, M., Scherer, K., 2010. Beyond arousal: valence and potency/control cues in the vocal expression of emotion. Journal of the Acoustical Society of America 128, 1322–1336.
Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R., 2007. Creating rapport with virtual agents. In: Lecture Notes in Artificial Intelligence; Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA), Paris, France, pp. 125–128.
Gravano, A., Benus, S., Hirschberg, J., German, E.S., Ward, G., 2008. The effect of prosody and semantic modality on the assessment of speaker certainty. In: Proceedings of 4th Speech Prosody Conference, Campinas, Brazil.
Gravano, A., Hirschberg, J., Beňuš, S., 2012. Affirmative cue words in task-oriented dialogue. Computational linguistics 38, 1–39.
Gustafson, J., Bell, L., Boye, J., Edlund, J., Wiren, M., 2002. Constraint manipulation and visualization in a multimodal dialogue system. In: Proceedings of the ISCA Workshop Multi-Modal Dialogue in Mobile Environments, Kloster, Irsee, Germany.
Gustafson, J., Heldner, M., Edlund, J., 2008. Potential benefits of human-like dialogue behaviour in the call routing domain. Perception and Interactive Technologies for Speech-Based Systems (PIT 2008). Springer, Berlin/Heidelber, pp. 240–251.
Gustafson, J., Neiberg, D., 2010. Prosodic cues to engagement in non-lexical response tokens in Swedish. In: The 5th Workshop on Disfluency in Spontaneous Speech and The 2nd International Symposium on Linguistic Patterns in Spontaneous Speech (DiSS-LPSS Joint Workshop 2010), Tokyo, Japan.
Heldner, M., Edlund, J., Hjalmarsson, A., Laskowski, K., 2011. Very short utterances and timing in turn-taking. In: 12th Annual Conference of the International Speech Communication Association (INTERSPEECH2011), Firenze, Italy.
Hirschberg, J., Litman, D., Swerts, M., 1999. Prosodic cues to recognition errors. In: Proceedings of the Automatic Speech Recognition and Understanding Workshop, pp. 349–352.
Hjalmarsson, A., 2008. Speaking without knowing what to say... or when to end. In: Proceedings of SIGDial 2008, Columbus, Ohio, USA.
Hjalmarsson, A., 2010. Human Interaction as a Model for Spoken Dialogue System Behaviour. Ph.D. thesis, Royal Institute of Technology (KTH).
House, D., 1990. Tonal Perception in Speech. Lund University Press, Lund.
Kendon, A., 1967. Some functions of gaze direction in social interaction. Acta Psychologica 26, 22–63.
Kopp, S., 2010. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Communication 52, 587–597.
Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., Stocksmeier, T., 2006. Modeling embodied feedback with virtual humans. In: ZiF Workshop, pp. 18–37.
Krahmer, E., Swerts, M., Theune, M., Weegels, M., 1999. The dual of denial: two uses of disconfirmations in dialogue and their prosodic correlates. Speech Communication 36, 133–145, ESCA Workshop on Dialogue and Prosody, September 1999.
Lai, C., 2009. Perceiving surprise on cue words: prosody and semantics interact on right and really. In: Proceedings of Interspeech'09, Brighton, UK.
Lai, C., 2010. What do you mean, you're uncertain?: the interpretation of cue words and rising intonation in dialogue. In: Proceedings of Interspeech 2010, Makuhari, Japan.
Larsson, S., 2002. Issue-based Dialogue Management. Ph.D. thesis, Goteborg University.
Laskowski, K., Jin, Q., Schultz, T., 2004. Crosscorrelation-based multispeaker speech activity detection. In: Proceedings of the 8th ISCA International Conference on Spoken Language Processing (INTERSPEECH2004), Jeju Island, South Korea, pp. 973–976.
Laukka, P., Audibert, N., Auberge, V., 2011. Exploring the determinants of the graded structure of vocal emotion expressions. Cognition Emotion, 37–41.
Levitt, E., 1964. The relationship between abilities to express emotional meanings vocally and facially. In: Davitz, J. (Ed.), The Communication of Emotional Meaning. McGraw-Hill, New York, pp. 87–100.
Liscombe, J., Venditti, J., Hirschberg, J., 2003. Classifying subject ratings of emotional speech using acoustic features. In: Proceedings of Eurospeech 2003.
Lloyd, S., 1982. Least squares quantization in pcm. IEEE Transactions on Information Theory 28, 129–137.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients: correction. Psychological Methods 1, 390.
Moravcsik, E.A., 1978. Reduplicative constructions. In: Greenberg, J.H. (Ed.), Universals of Human Language, . In: Word Structure, vol. 3. Stanford University Press, Stanford, pp. 297–334.
Neiberg, D., Gustafson, J., 2010. The prosody of Swedish conversational grunts. In: INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, pp. 2562–2565.
Neiberg, D., Gustafson, J., 2011a. A dual channel coupled decoder for fillers and feedback. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy.
Neiberg, D., Gustafson, J., 2011b. 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.
Neiberg, D., Gustafson, J., 2012a. Cues to perceived functions of acted and spontaneous feedback expressions. In: The Interdisciplinary Workshop on Feedback Behaviors in Dialog. Draft version.
Neiberg, D., Gustafson, J., 2012b. Exploring the implications for feedback of a neurocognitive theory of overlapped speech. In: The Interdisciplinary Workshop on Feedback Behaviors in Dialog.
Neiberg, D., Laukka, P., Elfenbein, H.A., 2011. Intra-, inter-, and cross-cultural classification of vocal affect. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, pp. 1581–1584.
Neiberg, D., Truong, K., 2011. Online detection of vocal listener responses with maximum latency constraints. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, pp. 5836–5839.
Nilsenova, M., 2006. Rises and Falls. Studies in the Semantics and Pragmatics of Intonation. Ph.D. thesis, University of Amsterdam.
Pammi, S., Schroder, M., 2009. Annotating meaning of listener vocalizations for speech synthesis. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009, pp. 1–6.
Payr, S., 2011. Social engagement with robots and agents: introduction. Applied Artificial Intelligence 25, 441–444.
Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., Kotz, S.A., 2009. Factors in the recognition of vocally expressed emotions: a comparison of four languages. Journal of Phonetics 37, 417–435.
Reese, B., 2007. Bias in Questions. Ph.D. thesis, University of Texas at Austin.
Reidsma, D., de Kok, I., Neiberg, D., Pammi, S., van Straalen, B., Truong, K., van Welbergen, H., 2011. Continuous interaction with a virtual human. Journal on Multimodal User Interfaces 4, 97–118.
Russell, J.A., 2003. Core affect and the psychological construction of emotion. Psychological Review 110, 145–172.
Sacks, H., Schegloff, E., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.
Sauter, D., Eisner, F., Ekman, P., Scott, S.K., 2010a. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proceedings of the National Academy of Sciences, 107.
Sauter, D.A., Eisner, F., Calder, A.J., Scott, S.K., 2010b. Perceptual cues in nonverbal vocal expressions of emotion. The Quarterly Journal of Experimental Psychology 63, 2251–2272.
Schegloff, E., 2000. Overlapping talk and the organization of turn-taking for conversation. Language in Society 29, 1–63.
Scherer, K., 1986. Vocal affect expression: a review and a model for future research. Psychological Bulletin 99, 143–165.
Scherer, K.R., 2009. The dynamic architecture of emotion: evidence for the component process model. Cognition and Emotion 23, 1307–1351.
Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck, T., 1991. Vocal cues in emotion encoding and decoding. Motivation and Emotion 15, 123–148.
Sigurd, B., 1984. Om jaså, bra, precis och andra returord. Språkvård, 3–8.
Skantze, G., 2007. Error Handling in Spoken Dialogue Systems – Managing Uncertainty, Grounding and Miscommunication. Ph.D. thesis, Royal Institue of TEchnology (KTH), Department of Speech, Music and Hearing.
Sloman, A., 2010. Requirements for artificial companions: it's harder than you think. In: Wilks, Y. (Ed.), Close Engagements With Artificial Companions: Key Social, Psychological, Ethical and Design Issues. John Benjamins.
Sokal, R.R., Rohlf, F.J., 1962. The comparison of dendrograms by objective methods. Taxon 11, 33–40.
Stocksmeier, T., Kopp, S., Gibbon, D., 2007. Synthesis of prosodic attitudinal variants in german backchannel ja. In: Proceedings of Interspeech, pp. 1290–1293.
Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C.V., Meteer, M., 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics 26, 339–373.
Stromqvist, S., Richthoff, U., 1999. Linguistic feedback, input and analysis in early language development. Journal of Pragmatics 31, 1245–1262.
Traum, D., 1994. A Computational Theory of Grounding in Natural Language Conversation. Ph.D. thesis, Computer Science Dept., U. Rochester, TR 545.
Wallers, Å., 2006. Minor Sounds of Major Importance – Prosodic Manipulation of Synthetic Backchannels in Swedish. Master's thesis, KTH Stockholm – School of Computer Science and Communication.
Ward, N., 2006. Non-lexical conversational sounds in american english. Pragmatics and Cognition 14, 129–182.
Ward, N., Tsukahara, W., 2000. Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32, 1177–1207.
Ward, N.G., Escalante, R., Bayyari, Y.A., Solorio, T., 2007. Learning to show you're listening. Computer Assisted Language Learning 20, 385–407.
Yngve, V.H., 1970. On getting a word in edgewise. Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, pp. 567–577.