Disambiguating the functions of conversational sounds with prosody: the case of 'yeah'

Khiet P. Truong and Dirk Heylen

Human Media Interaction Group, University of Twente, The Netherlands
{k.p.truong,d.k.j.heylen}@ewi.utwente.nl

Abstract

In this paper, we look at how prosody can be used to automatically distinguish between different dialogue act functions and how it determines degree of speaker incipiency. We focus on the different uses of 'yeah'. Firstly, we investigate ambiguous dialogue act functions of 'yeah': 'yeah' is most frequently used as a backchannel or an assessment. Secondly, we look at the degree of speakership incipiency of 'yeah': some 'yeah' items display a greater intent of the speaker to take the floor. Classification experiments with decision trees were performed to assess the role of prosody: we found that prosody indeed plays a role in disambiguating dialogue act functions and in determining degree of speaker incipiency of 'yeah'.

Index Terms: prosody, classification, dialogue act, speaker incipiency, backchannel

1. Introduction

'Yeah' to many people is a very common and seemingly simple word. However, in the research literature, 'yeah' is known as a rather complex, multifunctional word with ambiguous dialogue functions [1, 2, 3, 4, 5, 6]. 'Yeah' can for example function as a *backchannel*, i.e., a *continuer* [7]: by making head nods or short vocalizations such as 'mm-hmm' or 'yeah', the listener is showing the speaker that he/she is listening and that the speaker may carry on with the conversation. The speaker of the backchannel does *not* have the intention to take the conversational floor. In another context, 'yeah' may function as an *assessment* item: an assessment expresses an evaluation of something that was said previously. Assessments include, among other things, expressing agreement/disagreement, expressing approval/disapproval or expressing some attitude towards what has been said before. So, 'yeah' can be ambiguous in its dialogue act (DA) function: it is most frequently used as a backchannel or an assessment. But 'yeah' can also signal degree of *speaker incipiency* [1, 2, 5, 3, 8]. The degree of speaker incipiency is described in [1] as the level of the speaker's orientation toward taking the floor. Another description of speaker incipiency can be found in [2]: whether the speaker of the acknowledgment token continuous to speak immediately after enacting an acknowledgment token indicates the degree of speaker incipiency. So, some uses of 'yeah' will (pre-)signal the speaker's intention to take the floor whereas other uses of 'yeah' do not display this property.

The research presented in the current study can be placed in a broader context: we want to investigate how prosody is used to convey (conversational) meaning in vocalizations, and how we can use automatic analyses to detect these different meanings in vocalizations. In this paper, we narrow down our focus to the different uses of 'yeah'; we look at ways to disambiguate the different functions of 'yeah'. We do this by exploring *conversational state* features and prosody for classification. As classification algorithm, the decision tree C4.5 algorithm is used. With these methods and features, we first study how different DA functions of 'yeah' can be distinguished; is it used as a backchannel or as an assessment? Secondly, we look at how we can distinguish between low and high degree of speaker incipiency of 'yeah'. As possible application areas we target social signal processing and Embodied Conversational Agents (ECAs). For the analysis of social signals in conversation, classifying DAs and determining speakership incipiency can play an important role. Similarly, knowing how to display a greater degree of speaker incipiency can be useful for the design of ECAs.

2. Related work

Here, we will summarize some studies that are related to the disambiguation of functions of 'yeah' in conversation. Heylen and op den Akker [6] used conversational state features to distinguish between four DA functions of 'yeah': backchannel, assessment, stall and other. With a decision tree, an accuracy of 57% was achieved. No prosodic features were used. Shriberg et al. [9] did use prosodic features, i.e., duration, pause, and energy, to discriminate between agreements and backchannels, and achieved an accuracy of 69% using a decision tree. They found that agreements are longer in duration and have higher energy than backchannels. This is in contrast with what Benus et al. [10] found in their corpus; in their corpus, backchannels appeared to have a higher pitch and energy, and a greater pitch slope than affirmative words (agreements).

Regarding the degree of speaker incipiency of 'yeah', Drummond and Hopper [2] investigated which type of token - 'yeah', 'mm-hmm' or 'uh-huh' - is most likely to initiate further speech by its speaker. They found that 'yeah' is significantly more likely to display ongoing speakership than 'mm-hmm' or 'uh-huh'. They also found that only 55% of all 'yeah' tokens were freestanding, displaying no subsequent same-speaker speech, whereas in the case of 'mm-hmm' and 'uh-huh' this percentage is much higher. These observations support Jefferson's [3] claim that 'yeah' displays speakership incipiency while 'mm-hmm' displays passive speakership incipiency. Gardner [4, 5] investigated the pitch contour of 'yeah' and 'mm-hmm' tokens. He found that fall-rise pitch contours are more likely to be followed by talk from the immediately prior speaker, whereas falling contours are more commonly followed by same-speaker talk. This is in line with Jefferson's observation that 'yeah' typically carries a falling intonation contour, displaying speaker incipiency.

The results of the works described above all indicate that prosody indeed has additive value in distinguishing between DA functions and meanings of conversational sounds, as was also

suggested by the studies of Ward [11, 12]. However, few works have applied *automated* prosodic analysis to address the following classification tasks concerning 'yeah': 1) backchannel vs assessment (e.g. [9]), and 2) low vs high speaker incipiency. Hence, in the following Sections, we explain how we have addressed these two tasks.

3. Data - The AMI Corpus

The AMI Corpus is a multimodal dataset consisting of 100 hours of meeting recordings [13]. The meetings were transcribed ortographically and a large part of the corpus has been annotated for a wide range of behaviors such as dialogue acts, head movements, hand gestures, dominance and focus of attention. Each meeting has four speakers where each one of them plays a different role: a project manager, an interface expert, an industrial designer or a marketing expert. During the meetings, the participants discussed the development of a new remote control.

For the current study, we are using the DA annotations of 14 meetings[1]. We only consider turn-inital 'yeah's ($N = 1034$). It appears that 'yeah' is the most frequently used backchannel in our data (Table 1), and that it is also almost equally frequently used as an assessment (Table 2). In the dialog act manual of the AMI Corpus, attention is paid to this ambiguity: *'An ASSESS is any comment that expresses an evaluation, however tentative or incomplete, of something that the group is discussing [. . .] There are many different kinds of assessment; they include, among other things, accepting an offer, expressing agreement/disagreement or any opinion about some information that's been given, expressing uncertainty as to whether a suggestion is a good idea or not, evaluating actions by members of the group, such as drawings. [. . .] An ASSESS can be very short, like 'yeah' and 'ok'. It is important not to confuse this type of act with the class BACKCHANNEL, where the speaker is merely expressing, in the background, that they are following the conversation.'*

Lexical choices for the backchannel function	
Yeah	408
Mm/mm-hmm	370
Okay	127
vocalsound	57
Yes	36
Other	80
Total	1195

Table 1: *Counts of backchannels*

DA functions of 'yeah'	
Backchannel	408
Assess	352
Inform	119
Stall	52
Fragment	26
Other	77
Total	1034

Table 2: *Counts of turn-initial 'yeah'*

4. Method and Features

We present the conversational state features, the prosodic features, and the decision tree algorithm used in our experiments.

4.1. Conversational state features

Following the method carried out in [6], we make use of the notion of *conversational state*, representing an ensemble of the vocal activity of all participants to model the vocal interaction. Since there are four participants per meeting, a state is a 4-tuple <*a,b,c,d*> where a is the dialogue act performed by participant A etc. For example, <*Backchannel,0,0,Inform*> refers to a conversational state in which participant A was performing a Backchannel at the same time as participant D was performing an Inform. A conversation is in a particular state as long as no participant stops or starts speaking. Thus, a state change occurs each time some participant starts or stops speaking, i.e., starts or stops with his/her DA. We use the similar *conversational state* (CS) features as presented in [6], see Table 3. Ultimately, the goal is to predict events in time. As the features *continue* and *samespeaker* can only be known in retrospect, we prefer to use only past information, i.e. the features *lex* and *overlap* (referred to jointly as CS_past).

lex	0 if the DA consists of the word 'yeah' only, otherwise 1
continue	1 if the 'yeah'-speaker also speaks in the next conversational state, otherwise 0
same speaker	1 if one of the speakers, other than the 'yeah'-speaker, continues speaking in the next conversational state, otherwise 0
overlap	1 if there is more than one person speaking, otherwise 0
CS_all	lex+continue+samespeaker+overlap
CS_past	lex+overlap

Table 3: *CS features*

4.2. Prosodic features

The prosodic features can be divided into four sets (see Table 4). The first set of prosodic features consists of duration, mean pitch and mean intensity measured over the whole word 'yeah' (with Praat [14]). With the second set of features we tried to capture the falling/rising/fall-rise/rise-fall shape of the pitch contour which has shown to have distinctive properties [4, 5, 3]. This was done by fitting polynomial curves of degree 1 and 2 to the pitch values that were measured each 0.01s over the whole word 'yeah' and the second half of the word 'yeah'. The fourth set of features measured the absolute slope of pitch and the mean pitch over the second half of the word 'yeah'. All features were normalized to z-scores ($z = (x - \mu)/\sigma$) except for the coefficient features.

dur_z	duration over whole 'yeah' (z-score)
pitch_z	mean pitch over whole 'yeah' (z-score)
intens_z	mean intensity over whole 'yeah' (z-score)
absslopepitch_z	absolute slope of pitch over whole 'yeah' (z-score)
coef1, coef2	signs of 1st and 2nd degree polynomials, fitted on pitch values measured over whole 'yeah'
coef1_half, coef2_half	signs of 1st and 2nd degree polynomials, fitted on pitch values measured over 2nd half 'yeah'
pitch_half_z, absslopepitch_half_z	mean pitch and absolute slope of pitch measured over 2nd half of 'yeah' (z-score)

Table 4: *Four sets of prosodic features*

[1] ES2008a, IS1000a, IS1001a, IS1001b, IS1001c, IS1003b, IS1003d, IS1006b, IS1006d, IS1008a, IS1008b, IS1008c, IS1008d

4.3. Decision tree: C4.5 algorithm

Following [6] we used a C4.5 decision tree in WEKA [15] since the trees are relatively transparent and easy to understand. In short, at each node of the tree, the algorithm splits the dataset into subsets by testing attributes for certain values. The algorithm chooses the attribute that gives the highest normalized information gain as a result of the split, and classifies samples based on that split. It repeats this process recursively in each subset obtained until all the samples are classified or until splitting does not have additive value.

5. Experimental setup

We first present the experimental setup for the classification experiments. All classification experiments were carried out in WEKA [15] with a 10-fold cross-validation. As data, only turn-initial 'yeah's were considered. Out of the total of 1034 'yeah' samples, 17 were discarded since no pitch could be measured at all in these samples. As evaluation performance measures, the averaged Area-Under-Roc (AUC) and (macro) average F_1-measure were used.

5.1. Classifying the DA function of 'yeah'

We were interested in how well the features under investigation could separate the different DA functions of 'yeah'. Since **bck** (backchannel) and **ass** (assessment) form the largest classes (Table 2), one of the experiments considers **bck vs ass vs oth** where **oth** contain the rest of 'yeah' that were not **bck** or **ass**. In addition, we looked at the binary decision **bck vs ass**. In total, we experimented with 402 **bck**, 346 **ass**, and 269 **oth** samples. All combinations of sets of features (see Table 3 and 4) were evaluated, and the first four prosodic features as seen in Table 4 were also individually combined with all other possible combinations (since in pilot experiments, these proved to be strong individual features).

5.2. 'Yeah' as a speaker incipiency marker

It has been suggested that 'yeah' also signals a certain level of speaker incipiency that is at least higher than the continuer 'mm/mm-hmm', e.g. [1, 3]. Some of the 'yeah's will function as true continuers and have low speaker incipiency, whereas other 'yeah's may function as (pre-)signals to turn-taking, i.e., high speaker incipiency, and may signal that the speaker is about to take the floor or the turn. Using the CS features and the prosodic features, we explored whether the speaker incipiency of 'yeah' could be predicted based on these features. In order to have positive and negative examples of speaker incipiency we defined a *working definition* of speaker incipiency: the number of *conversational states* that have passed till the current speaker starts a new turn determines low or high speaker incipiency. The length of this new turn needs to contain at least 10 words, otherwise it is not considered a 'new turn'. For the current classification experiments, we consider a number of CSs between 1 and 7 (which corresponds to an average time of 9.6s, sd of 8.6s) as highly speaker incipient ($+$**SpIn**) and a number of CSs larger than 30 (which corresponds to an average time of 41.1s, sd of 18.6s) as low speaker incipient ($-$**SpIn**), everything in between was discarded. As a result, we obtained 303 $+$**SpIn** samples of high and 305 $-$**SpIn** samples of low speaker incipiency.

6. Results

In addition to the results of the classification experiments, we will also take a closer look at the prosody of 'yeah' in relation to its DA function and speaker incipiency.

6.1. Classifying the DA function of 'yeah'

Using all CS features similar to [6], an AUC of 0.69 was obtained in the **bck vs ass vs oth** experiments. This performance dropped to 0.64 when the retrospective features, *overlap+samespeaker*, were removed, see Table 5. Adding prosody to this feature set significantly improved AUC to 0.67 (paired T-test, $p<0.05$) . A similar effect can be found for the **bck vs ass** experiments: adding dur_z to the CS_past feature set signficantly improved AUC from 0.65 to 0.70 (paired T-test, $p<0.05$).

Experiment	Features	AUC	F_1
bck vs ass vs oth	CS_all	0.69	0.52
bck vs ass vs oth	CS_past	0.64	0.43
bck vs ass vs oth	CS_past+dur_z	0.67	0.48
bck vs ass	CS_all	0.72	0.70
bck vs ass	CS_past	0.65	0.65
bck vs ass	CS_past+dur_z	0.70	0.68

Table 5: *Results of the decision tree classification experiments: DA of 'yeah'*

We found that the best results obtained always included the dur_z feature. In Table 6 some prosodic features that are (almost) significant at $p<0.05$ different from each other for the **bck**, **ass**, and **oth** classes are shown. It seems that backchannels usually have a shorter duration and lower intensity than assessments and other DAs.

	pvalue	**bck**	**ass**	**oth**
dur_z	0.0235	**-0.052**	**0.073**	-0.016
intens_z	0.0366	**-0.096**	0.026	**0.111**
meanpitch_half_z	0.1102	-0.083	0.079	0.110

Table 6: *Means of significant prosodic features for the analysis of the DA function of 'yeah' (**bold** means that the distributions are signficantly different from each other, Kruskal-Wallis test with post-hoc Tukey-Kramer $p<0.05$)*

6.2. 'Yeah' as a speaker incipiency marker

We also explored the use of 'yeah' as a speaker incipiency marker and looked at whether certain features could predict low or high speaker incipiency with 'yeah'. Table 7 shows that there are indeed attributes in the prosody and vocal activity of the speaker that can predict, better than chance, whether 'yeah' was used merely as a continuer or whether it was used to signal that he/she is about to make a speaker turn. The addition of prosodic features improved AUC from 0.55 to 0.58 (Table 7), although this increase was not statistically significant ($p<0.05$). However, dur_z appeared to be a useful prosodic feature: a high speaker incipient 'yeah' generally has a longer duration than a low speaker incipient 'yeah'. Since speaker incipiency might be related to the DA type of 'yeah', we looked at the distribution of DA types in the data regarding $+$**SpIn** and $-$**SpIn**. As

expected, **ass** and **bck** were the largest classes of DA types for +**SpIn** and −**SpIn** respectively, but other DA types also played a role (which also became clear when we calculated the amount of overlap, 51%, between the two datasets used in **bck vs ass** and +**SpIn vs** −**SpIn**).

Experiment	Features	AUC	F_1
+**SpIn vs** −**SpIn**	CS_all	0.65	0.63
+**SpIn vs** −**SpIn**	CS_past	0.55	0.52
+**SpIn vs** −**SpIn**	CS_past + dur_z + intens_z + coef1_half + coef2_half + abss-lopepitch_z	0.58	0.56

Table 7: *Results of the decision tree classification experiments: speaker incipiency of 'yeah'*

In comparison with other frequently occurring backchannels such as 'mm' and 'okay', 'yeah' is the one that shows the highest level of speaker incipiency ([2, 3, 1, 4]). We assessed this observation by counting the number of CSs that have passed until the same speaker starts a new speaker turn (a speaker turn has to be comprised of at least 10 words), and by counting the number of words in the first next utterance that the speaker makes (without the constraint that this has to be a speaker turn), see Table 8. The counts show that people start a new speaker turn significantly ($p<0.05$) sooner after they have said 'yeah' than 'mm'. In addition, 'yeah' is generally followed by longer turns uttered by the same speaker than in the case of 'mm'. Finally, the *continue* feature indicates that 'yeah' is more often followed by same-speaker speech.

	'yeah'	'mm'	'okay'
averaged number of CSs until the speaker starts a new turn again	31	49	35
number of words in the next CS	5.0	3.7	4.7
% of CS feature *continue*= 1	61%	42%	59%

Table 8: *Differences in degree of speaker incipiency of 'yeah', 'mm/mm-hmm', and 'okay'*

7. Conclusions and Discussion

We have disambiguated different functions of 'yeah' with conversational state features and prosody. The first experiments concerned the distinction between backchannels, assessments and other DAs. Averaged AUCs between 0.67 and 0.70 were achieved. Adding prosodic features significantly improved performance. We found that 'yeah' used as an assessment is generally longer in duration than a backchannel 'yeah' (similar to [9]). In the second sets of experiments, we looked at whether the degree of speaker incipiency of 'yeah' could be predicted with conversational state features and prosody. We obtained an averaged AUC of 0.55 which was improved to 0.58 when prosodic features were added (although this addition was not significant). In addition, we confirmed that 'yeah' in general has a greater degree of speaker incipiency than 'mm/mm-hmm'.

In general, the effect of prosody was smaller than expected. We expected to find more predictive value in the shape of the intonation contour as this was reported as a distinctive property in [4, 5, 10, 3]. Some of the reasons why the effect of prosody is less apparent than expected could be that our prosodic analyses were all performed automatically, and, although the AMI word-level transcriptions provided were obtained manually, we experienced in practice that sometimess the word boundaries were off. In addition, it should be noted that we adopted a 'mechanical' definition of speaker incipiency that is subject to discussion. For future research, we suggest to look more closely at other ways to predict speaker incipiency; are there other (multimodal) markers that display a greater degree of speaker incipiency? Also, in designing ECAs, we can take into account the findings discussed here and use (or test) 'yeah' as a speaker incipient signal.

8. Acknowledgements

This research has been supported by the European Community's 7th Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet).

9. References

[1] K. Drummond and R. Hopper, "Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency," *Research on Language and Social Interaction*, vol. 26, pp. 157–177, 1993.

[2] ——, "Some Uses of Yeah," *Research on Language and Social Interaction*, vol. 26, no. 2, pp. 203–212, 1993.

[3] G. Jefferson, "Notes on a Systematic Deployment of the Acknowledgement Tokens "Yeah" and "Mm hm"," in *Tilburg Papers in Language and Literature*, 1984.

[4] R. Gardner, "The Conversation Object Mm: A Weak and Variable Acknowledging Token," *Research on Language and Social Interaction*, vol. 30, no. 2, pp. 131–156, 1997.

[5] ——, "Between Speaking and Listening: The Vocalisation of Understandings," *Applied Linguistics*, vol. 19, pp. 204–224, 1998.

[6] D. Heylen and R. op den Akker, "Computing Backchannel Distributions in Multi-Party Conversations," in *Proceedings of the ACL Workshop on Embodied Language Processing*, 2007, pp. 17–24.

[7] E. A. Schegloff, "Discourse as an Interactional Achievement: Some Uses of 'uh huh' and Other Things That Come Between Sentences," in *Georgetown University Roundtable on Languages and Linguistics 1981 - Analyzing Discourse: Text and Talk*, D. Tannen, Ed. Georgetown University Press, 1982, pp. 71–93.

[8] D. H. Zimmerman, "Acknowledgment Tokens and Speakership Incipiency Revisited," *Research on Language and Social Interaction*, vol. 26, no. 2, pp. 179–194, 1993.

[9] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. van Ess-Dykema, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[10] S. Benus, A. Gravano, and J. Hirschberg, "The Prosody of Backchannels in American English," in *Proceedings of ICPhS2007*, 2007, pp. 1065–1068.

[11] N. Ward, "Pragmatic Functions of Prosodic Features in Non-Lexical Utterances," in *Proceedings of Speech Prosody 2004*, 2004, pp. 325–328.

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

[13] J. Carletta, "Unleashing the Killer Corpus: Experiences in Creating the Multi-Everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, pp. 181–190, 2007.

[14] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer (Version 5.1.07)," 2009. [Online]. Available: http://www.praat.org

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.