Positioning and acoustics of French feedback items

Jan Gorisch[1]*, Laurent Prévot*[1]

[1]Aix-Marseille Université, CNRS, Laboratoire Parole et Langage, France

jan.gorisch@lpl-aix.fr, laurent.prevot@lpl-aix.fr

Abstract

Verbal feedback is regularly used in dialogue by interacting participants. It is debatable whether its positioning and its acoustic makeup is random or if there are certain regularities on the contextual use and the prosodic form of feedback. This study investigates how the French feedback items "oui" and "ouais" are employed by participants in corpora of conversational interactions and of task-oriented dialogue. It is suggested that certain contextual positioning, e.g. feedback in isolation vs. feedback in overlap, co-occurs with specific acoustic-prosodic properties, e.g. similarities between pitch contour of the feedback item and the contour of the interlocutor.

Index Terms: corpus, conversation, task-oriented dialogue, pitch contour, prosodic similarity

1. Introduction

English feedback items such as "yeah", "okay" and "uh-huh" have been described as *back-channels* [1], *response tokens* [2] or in general as *feedback* [3, 4]. Beyond their definitions it is generally agreed, that they can express different communicative functions. Contextual cues indicate what action is performed at what time. For simple back-channels for example it is assumed that they merely indicate that an interlocutor is attending to the other interlocutor without the attempt to grasp the floor. Schegloff termed this a *continuer*, as it makes the priorly speaking interlocutor continue speaking [5]. These items do not occur at random places, but at specific points in the participant's talk, namely at Transition Relevance Places (TRPs), where the turn of the prior speaker's turn comes to a potential point of completion [6]. At such a place, other social actions can be performed, such as *assessments* ("that's good", "wow"). This has been demonstrated qualitatively in the framework of Conversation Analysis (CA).

Decision tree modeling has been used in interactional phonetics (IP; [7]) work in order to investigate positioning and prosodic features related to overlap competitiveness [8]. It has been shown that competitive and non-competitive overlaps can be distinguished based on overlap *positioning* features with regard to context, e.g. syntactic turn completion features of prior talk, recycling of words, the number of words and syllables in overlap, *pausing* features, e.g. pause position, duration, frequency compared to the number of words, and *prosodic* features, e.g. similarity of the F0-contour in overlap compared to contours in the clear, F0-mean and F0-mean compared to the F0-mean of the interlocutor's pre-overlap talk.

Such developments in research using large collections of feedback items from corpora require annotations of turns into categories. They are either performed by naïf coders using annotation guidelines for categories such as backchannels [9] or they are performed by experts using CA for classes such as comptetitive vs. non-competitive overlap [8] and alignments vs. non-alignments [10].

The risk of manual annotation of categories is that some aspects of that procedure–may they be involved directly in the annotation guidelines or implicitely in the methodology used for the annotation–can re-appear in the analysis of the data afterwards. For example the explanation of the CA-informed categorisation into competitive and non-competitive overlaps [8] makes use of the syntactic cue of completeness. If the prior speaker's turn is syntactically incomplete at the start of the overlap, the likelihood of annotating the overlap as competitive increases (of course, apart from other factors). In the analysis phase, it can happen that a feature is used that is very similar to that cue. For example the feature *turn completion* [8, p.733] is part of the strong feature set that distinguishes the competitiveness even better than the prosodic feature set. This shows how difficult it is to find a balance between the ingredients taken form the various disciplines that can inform us about turn-taking.

From a data-focused perspective such as data mining using machine learning techniques that have no qualitative annotations available, it can be anticipated that automatic clustering (for example) of a huge amount of features and instances, as Kurtic et al. [8] have available, separates the data into classes that resemble the interactional categories previously postulated.

It is hypothesised that the use of "oui" and "ouais" is more complex than described in the Nouveau Petit Robert [11] that classifies "oui" as affirmative adverb and "ouais" as interjection that expresses surprise and as a colloquial expression for "oui". The time-consuming task of annotating feedback items according to interactional categories, i.e. communicative functions or social actions, is postponed in the present work and substituted by an unsupervised classification based on parameters that can be extracted from the transcriptions, such as overlap (in overlap or not, overlap duration), placement (isolated, initial, final, other), pause duration (of the speaker uttering the feedback item and the interlocutor before and after the feedback) and speech duration (before and after). These fall into the basket of positional features. Prosodic features are based on the fundamental frequency (F0 slope, height, min., max., span) and pitch contour similarity (between the feedback item and the last item of the prior turn of the interlocutor).

2. Material

2.1. Corpora

Two collections are investigated, the Conversation Interaction Data (CID) corpus and the French Map Task corpus. The CID [12] consists of free conversations between work col-

leagues. Two participants were given the instruction to talk about "strange things" and were recorded during one hour using headset microphones. The corpus contains eight pairs of that kind of interaction. Three pairs have been recorded additionally with a video camera.

The Aix Map Task corpus [13] consists of task-oriented dialogue. Two interactional participants reconstruct a path that is drawn on the map of one participant (the giver) onto the map of the other participant (the follower). Two conditions have been recorded. In the audio-only condition, the participants were specially separated and could not see each other. The communication occurred merely across the acoustic channel. In the face-to-face condition, participants were seated in front of each other. Additionally to the acoustic modality, also the visual modality was therefore available. Both conditions are recorded via headset microphones and in the face-to-face condition with individual cameras for each participant. In each condition, four pairs of participants were recorded. The duration of the interactions varied according to the maps and the participants' performance and took mainly between 30 and 40 minutes (for 8 maps).

2.2. Lexical items

The French feedback items "oui" [wi] ("yes") and "ouais" [wɜ] ("yeah") are taken for this study because their phonetic structure is relatively simple. Both are prototypically produced in one syllable and contain merely voiced segments. For these reasons, other potential feedback items were excluded, such as "d'accord" [dakɔʀ], "voilà" [vwala], "mm-hmm" [mhm].

3. Method

The whole data set (6344 instances) contains 5093 instances of "ouais" and 1251 "oui". Most of them (3748) stem from the CID corpus, 1394 from the audio-only condition and 1202 from the face-to-face condition of the Map Task corpus. 30 distinct speakers are involved. Two types of features are recorded: positional and prosodic ones.

3.1. Position features

The position of the feedback item is considered according to pausing, i.e. when the speaker of the feedback item or the interlocutor do not speak, and according to the overlap of both speakers.

pb the *pause before* the feedback item

pa the *pause after* the feedback item

opb the *other* speaker's *pause before* the feedback item

opa the *other* speaker's *pause after* the feedback item

dur the *duration* of the feedback item

do the *duration of overlap* between feedback item and interlocutor

ndo the *normalised overlap duration*: do / dur

osb the *overlapping speech before*

osa the *overlapping speech after*

Depending on the pausing and overlap properties of the talk (IPU = Inter Pausal Unit) embedding the feedback item, an overall positioning feature **pos** indicates whether the feedback item is produced in isolation *iso*, IPU initial *ini*, IPU final *fin* or other *oth*. This is illustrated in Figure 1. The features related to pausing are illustrated in Figure 2.

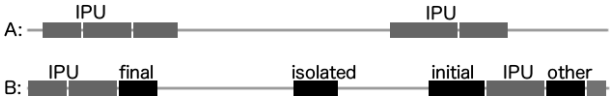

Figure 1: Positions of feedback items within an IPU: final, isolated, initial and other.

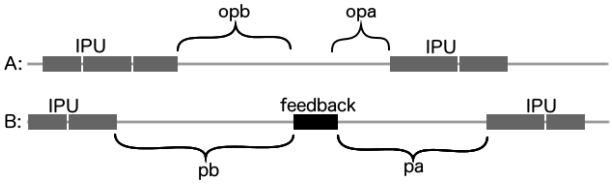

Figure 2: Pause features of feedback-items in isolation: pb and pa (pause before and after), opb and opa (overlap pause before and after).

3.2. Acoustic-prosodic Features

The following prosodic features have been extracted for the contour of the feedback item and the last preceding word of the interlocutor:

slope the F0 slope

f0min the minimum F0 value

f0max the maximum F0 value

f0stdev the standard deviation of all F0 values

span the difference between the f0max and f0min

height the mean of the F0 values

An additional feature was used to indicate the similarity between pitch contours of the feedback item and the preceding word of the interlocutor (**SimScore**) following [14, **?**]. The target contour is the one of the feedback item and the domain contour is determined by the following factors: The domain is a single word in the interlocutor's speech that precedes or overlaps the feedback item. The scope for the search of that word is limited to three seconds. If the interlocutor's pause before the item exceeds this time limit, the similarity score is not computed for that item. The domain word is allowed to overlap the beginning of the feedback item, however, it may not exceed it.

3.3. Unsupervised clustering

K-means clustering with the Euclidean Distance was used as implemented by Arthur and Vissilvitskii [15]. It is provided by the Weka toolkit[1]. The aim is to partition the 6344 observations into 2 sets so that the within-cluster sum of squared errors minimises. The feature *pos* is ignored by the algorithm as it is modeled by the pausing features and therefore redundant.

4. Results

Table 1 shows the results from the SimpleKMeans clustering experiment. From the 6344 instances, 4943 (78%) fall into cluster 1 and 1401 (22%) fall into cluster 2. After 7 iterations, the within cluster sum of squared errors is at 6830.2. The table displays the cluster centroids for each attribute. Some positional features don't differ largely between clusters with regard to the

[1] available from http://www.cs.waikato.ac.nz/ml/weka/.

centroid of the cluster on the Full Data set. Those are the positional features associated with the feedback item speaker (pause before (*pb*) and after (*pa*) the feedback item) and the features associated with overlapping speech (duration and normalised duration of overlap (*do* and *ndo*) and speech of the feedback item speaker before (*osb*) and after (*osa*) the feedback item that is overlapped by the other speaker). (The values of 0.0 in *pb* and *pa* come from the feedback items (3226 instances) that are not isolated, i.e. preceded or followed by further talk.) However, the pause of the other speaker that comes before the feedback item (*opb*) and after (*opa*) seem to differ across cluster centroids. While the overall centroid of *opb* is at 0.550, cluster 1's centroid is above at 0.639 and cluster 2's centroid is below at 0.235. The same relationship can be observed for feature *opa*, where Cl_1 >All >Cl_2. Similar relationships can be found in the prosodic features. Although the slope is relatively flat (around 0), cluster centroids diverge: Cl_1(*slope*) <All <Cl_2(*slope*). Similar tendencies occur for all other prosodic features: Cl_1(*feature*) <All <Cl_2(*feature*). The prosodic features, including pitch contour similarity (*SimScore*) seem to depend on each other. Lower values in *slope*, *f0min*, *f0max*, *f0stdev*, *span* and *height* fall into the cluster Cl_2 with relatively low pitch contour similarity. High values in these features fall into Cl_1 with relatively high pitch contour similarity.

The results suggest a relationship between the pausing behaviour of the interlocutor and the prosodic features of the feedback item, the last word of the interlocutor before the feedback item and the pitch contour similarity of the adjacent words. This relationship is: a shorter pause of the prior speaker before and after the feedback item (*opb* and *opa* is associated with a lower pitch contour similarity and lower values in the other prosodic features, while a longer pause is associated with higher pitch contour similarity and higher values in the other prosodic features.

5. Discussion

The interpretation of results from a clustering experiment largely depend on the underlying data [16]. The data are drawn from conversational talk and task-oriented dialogue and are therefore an extreme challenge for the analysis of prosody. The orthographic transcription, semi-automatic alignment and the signal itself involving unavoidable cross-talk have an influence on every subsequent analysis step. With that in mind an example should be described here that illustrates the positioning of the feedback item, the f0 contours and the pitch contour similarity score.

The example is drawn from the CID corpus, constituting talk-in-interaction from two female participants (AB and CM). Figure 3 illustrates the F0 contours including the feedback item and the last three seconds before it. AB has currently the main speaker role, narrating a story, while CM is giving feedback in form of many "ouais". The "ouais" that has been picked here occurs at second 3. It is preceded and followed by further talk from the same speaker (there is no pause before or after). It does not overlap the prior talk from the interlocutor ("... je crois"/*"I think"*). The last word is therefore the "crois" that represents the domain to which the feedback item is compared to prosodically.

The F0-contour of the domain speaker (AB) in the word "crois" falls steeply at it's beginning, i.e. at the transition from the voiceless sounds [kʀ] to the bilabial opening of the vowel [wa] that contains the rest of the contour near the level of the speaker's mid F0-range. The F0-contour of the feedback item (CM) fluctuates slightly, but is equally produced close to the mid range of that speaker.

Following [14], a pairwise comparison of all F0-values from both F0-contours creates a similarity matrix (see Figure 4). Subsequent Dynamic Time Warping through the similarity matrix finds the best alignment path (red line) with its associated quality score. In this example, the normalized *SimScore* is 0.89 on a scale from 0 to 1.

Examples, where the comparison of F0-contours does not work that neatly can be imagined when the other "ouais" from speaker CM in Figure 3 are considered. The F0 values might be extremely sparse, as for example the "ouais" between second 2 and 2.5. Or the F0 contour of the domain might be missing as for the first half of the interlocutor's talk between second 0 and 2.

Table 1: Results for k-means clustering of the current data set. For each attribute the corresponding cluster centroid is indicated for the Full Data, Cluster 1 and Cluster 2.

Attribute (Feature)	Full Data (6344)	Cluster 1 (4943)	Cluster 2 (1401)
pa	0.0	0.0	0.0
pb	0.0	0.0	0.0
do	0.149	0.144	0.168
ndo	0.514	0.499	0.568
osb	1.281	1.247	1.399
osa	0.941	0.920	1.015
opb	0.550	0.639	0.235
opa	1.066	1.137	0.815
slope	-0.007	-0.003	-0.019
slopeInterl	-0.003	-0.002	-0.009
f0min	-0.602	-0.599	-0.615
f0minInterl	-0.605	-0.623	-0.542
f0max	0.315	0.366	0.133
f0maxInterl	0.392	0.419	0.297
f0stdev	0.279	0.290	0.239
f0stdevInterl	0.285	0.295	0.250
span	0.917	0.965	0.748
spanInterl	0.997	1.041	0.838
height	-0.091	-0.051	-0.231
heightInterl	-0.019	-0.010	-0.053
SimScore	0.295	0.372	0.024

6. Conclusions and future work

The experiment described in this paper investigated the possibility to automatically cluster instances of French feedback items restricted to "oui" and "ouais" according to positional and prosodic-acoustic features.

Assuming that the French feedback items "ouais" and "oui" are in their majority employed by interactional participants in order to perform aligning and affiliative rather than non-aligning and disaffiliative actions [17] [18], it can be expected that they are on average produced with high prosodic similarity [19] [10], also shown to signal entrainment [20]. Furthermore, a difference in the prosodic form according to the positioning of feedback items can be expected [8].

The results suggest that these feedback items are highly variable and it is difficult to cluster them into classes that can be distinguished and interpreted according to placement, prosody

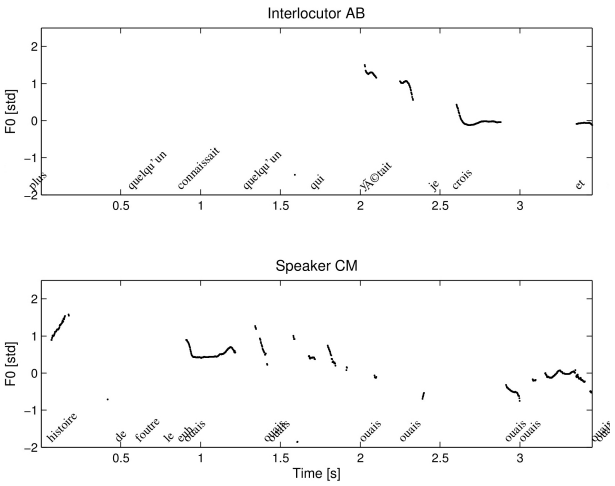

Figure 3: F0 contours of the feedback item speaker (bottom) and interlocutor (top. The ordinate indicates the F0 in standard deviations according to the speaker's mid range.

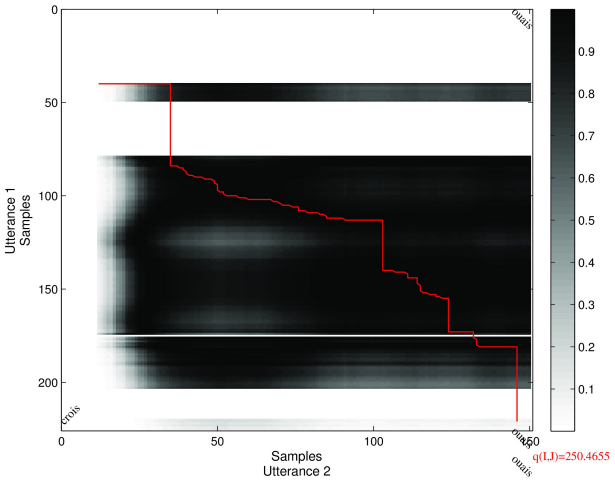

Figure 4: Similarity Matrix with the alignment path of the F0 contours.

and pitch contour similarity [**?**]. The clusters do not follow speaker characteristics or dependencies on the interactional situation (conversation vs. task-oriented dialogue). However, the clustering experiment has shown that the feedback items with specific prosodic characteristics are placed according to the pausing structure of the interlocutor. If the pause of the interlocutor preceding the feedback item is short, the pitch contour similarity is lower than if the pause is long. This means that feedback items cannot be placed anywhere with any prosodic makeup.

Laurent: Acturally, I need to double check the original csv I provided because when I looked at Léo's stats on filled pauses, I realized that there was a problem with one of my csv... I hope it is not the one I sent you as well...

The results of this study can further be improved, as the selection of features are partially redundant and partially incomplete. For example, the duration of overlapping speech of the interlocutor preceding and following the feedback item of the target speaker *osb* and *osa* was recorded while the overall duration of talk before the feedback item of both speakers was neglected. A feature is missing that measures the proportion of speech (beginning or end of the feedback item) that overlaps with–or that is overlapped by–speech from the interlocutor. That is important from a point of view of competitiveness of overlapping talk [8] that can have an influence on the choice of prosodic contour by the speaker of the feedback item.

A potential source for error comes from the orthographic transcription of a part of the Map Task corpus that has not been corrected manually yet after automatic segmentation into inter-pausal-units and alignment of the transcriptions. The transcription and alignment tool SPPAS [21] had thereby to cope with unavoidable crosstalk, which had as consequence erroneous segmentation. The cleaning of IPU segmentation is part of current work.

Future work involves the annotation of all feedback items–also the ones mentioned before, such as "d'accord", "voilà", "ok"–into categories of communicative functions as outlined in [4]. Such categories can then serve as a ground truth, to which models of classifiers can be compared.

Despite the sparseness of F0 readings and other problems encountered in this study that may occur in conversational talk and task-oriented dialogue, it was shown that the analysis of such data can bring interesting insights into the organisation of talk-in-interaction and the use of prosodic and positional resources for that organisation.

7. Acknowledgements

This research is funded by the ANR project "Conversational Feedback" (Grant Number: ANR-12-JCJC-JSH2-006-01). And the Erasmus Mundus Multi Exchange Program. We are grateful for the support of the transcribers.

8. References

[1] V. Yngve, "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, 1970.

[2] R. Gardner, *When listeners talk: Response tokens and listener stance*. Amsterdam; Philadelphia: John Benjamins, 2001.

[3] H. Bunt, "Context and dialogue control," *Think Quarterly*, vol. 3, no. 1, pp. 19–31, 1994.

[4] L. Prevot and R. Bertrand, "CoFee—toward a multidimensional analysis of conversational feedback, the case of French language," in *Proceedings of the Workshop on Feedback Behaviors in Dialog*, 2012.

[5] 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.

[6] H. Sacks, E. A. Schefloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[7] J. Local, "Phonetic detail and the organisation of talk-in-interaction," in *Proceedings of the 16th ICPhS*, 2007.

[8] E. Kurtic, G. J. Brown, and B. Wells, "Resources for turn competition in overlapping talk," *Speech Communication*, vol. 55, pp. 721–743, 2013.

[9] A. Gravano and J. Hirschberg, "Affirmative cue words in task-oriented dialogue," *Computational Linguistics*, vol. 38, no. 1, pp. 1–39, 2012.

[10] J. Gorisch, B. Wells, and G. J. Brown, "Pitch contour matching and interactional alignment across turns: an acoustic investigation," *Language and Speech*, vol. 55, no. 1, pp. 57–76, 2012.

[11] J. Rey-Debove and A. Rey, *Le nouveau petit Robert. Dictionnaire alphabétique et analogique de la langue française*. Montreal, Canada: Dicorobert Inc., 1996.

[12] R. Bertrand, P. Blache, R. Espesser, G. Ferre, C. Meunier, B. Priego-Valverde, and S. Rauzy, "Le CID - Corpus of Interactional Data-annotation et exploitation multimodale de parole conversationnelle," *Traitement automatique des langues (TAL)*, vol. 49, no. 3, pp. 105–134, 2008.

[13] J. Gorisch, C. Astésano, E. G. Bard, B. Bigi, and L. Prévot, "Aix map task corpus: The french multimodal corpus of task-oriented dialogue," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014.

[14] J. Gorisch, "Matching across turns in talk-in-interaction: the role of prosody and gesture," Ph.D. dissertation, The University of Sheffield, 2012.

[15] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of carefull seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[16] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2011.

[17] D. Barth-Weingarten, "Double sayings of ja–more observations on their phonetic form and alignment function," *Research on Language and Social Interaction*, vol. 44, no. 2, pp. 157–185, 2011.

[18] T. Stivers, "Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation," *Research on Language and Social Interaction*, vol. 41, no. 1, pp. 31–57, 2011.

[19] B. Szczepek Reed, "Beyond the particular: Prosody and the coordination of action," *Language and Speech*, vol. 55, no. 1, pp. 13–34, 2012.

[20] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of NAACL 2012*, Montreal, Canada, 2012.

[21] B. Bigi, "SPPAS: A tool for the phonetic segmentations of speech," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012, pp. 1748–1755.