Towards the generation of dialogue acts in socio-affective ECAs:
A corpus-based prosodic analysis

Rachel Bawden*[1], Chloé Clavel†[1], and Frédéric Landragin‡[2]

[1]Institut Télécom, Télécom-ParisTech, CNRS LTCI
[2]Lattice-CNRS, ENS & University Sorbonne Nouvelle

Authors' draft

Abstract

We present a corpus-based prosodic analysis with the aim of uncovering the relationship between dialogue acts, personality and prosody in view to providing guidelines for the ECA Greta's text-to-speech system. The corpus used is the SEMAINE corpus, featuring four different personalities, further annotated for dialogue acts and prosodic features. In order to show the importance of the choice of dialogue act taxonomy, two different taxonomies were used, the first corresponding to Searle's taxonomy of speech acts and the second, inspired by Bunt's DIT++, including a division of directive acts into finer categories. Our results show that finer-grained distinctions are important when choosing a taxonomy. We also show with some preliminary results that the prosodic correlates of dialogue acts are not always as cited in the literature and prove more complex and variable. By studying the realisation of different directive acts, we also observe differences in the communicative strategies of the ECA depending on personality, in view to providing input to a speech system.

1 Introduction

Embodied conversational agents (henceforth ECAs) are animated virtual characters capable of engaging in conversation with a human user. Their purpose is to provide realistic and natural communicative behaviour, usually in the context of a particular task, such as providing support for the user. The task of animating such an agent therefore regroups a number of different research topics including speech synthesis, speech recognition, motion capture, motion generation and the recognition and generation of emotion. Engaging in conversation involves multimodal communicative behaviour: speech, gesture and facial expressions, and the appropriate coordination of these three modalities is crucial for successful and natural conversation. ECAs can

*rachel.bawden@keble.oxon.org
†chloe.clavel@telecom-paristech.fr
‡frederic.landragin@ens.fr

be used in various applications, such as the role of an assistant on sales sites (Suignard, 2010) or of a tutor in Serious Games (Klatt *et al.*, 2011). For example, the EU TARDIS project aims to develop a serious game using ECAs to train teenagers for job interviews (Anderson *et al.*, 2013), and the A1:1 French project (Campano *et al.*, 2015) seeks to develop interaction between a virtual agent and museum visitors, with a focus on encouraging user engagement.

Current architectures for virtual agents mainly use pipelined modules to separate and integrate different aspects of communication, and many focus on the generation of multimodal behaviour (speech, gesture and facial expressions) as well as adding socio-effective components to introduce personality or emotion into the agent's behaviour. There can be an advanced handling of gesture, with timecodes used to synchronise gesture and speech, and different communicative behaviour associated with individual personalities. Formats such as FML-APML (the Affective Presentation Markup Language (De Carolis *et al.*, 2004) based on the Functional Markup Language) can be used to encode multimodal communicative intentions, with manually written prosodic settings associated to regions of speech via prosodic xml tags. Prosody is one of the aspects of communicative behaviour that appears to be amongst those most influenced by the other various communicative factors (dialogue act, personal aims, emotion, personality and speaker-addressee relationship to name just a few). However at present, prosody must either be hardcoded in order to be processed by a speech synthesis system or be inherent to the corpus used for text-to-speech (TTS) systems and in this case is far less configurable. We consider that it is important to maintain control over the prosodic settings, and what is generally lacking is the automatic generation of prosodic parameters based on different intentions and various other factors. Automatic generation of prosody requires understanding the link from the other communicative factors to prosody and also how these factors might interact, something that is poorly understood, especially in the context of generating believable speech.

This study therefore aims to explore the relationship between two aspects, the type of dialogue act and personality, and prosodic generation in view to providing guidelines for mapping these features to prosody for a configurable TTS synthesis system. We will study the interaction through a corpus study of the SEMAINE corpus (McKeown *et al.*, 2012), designed to provide emotional and spontaneous dialogue, and annotated for dialogue acts and prosodic features. By comparing two different dialogue act taxonomies, we will assess the degree to which the distinctions made can be useful for identifying definable prosodic correlates of dialogue acts. We will also study the relationship between the different personalities present in the SEMAINE corpus and their prosodic realisations to better understand how prosody can be designed to emulate them.

The architecture that will benefit from the present analysis is the Greta architecture (Bevacqua *et al.*, 2010), a modular system capable of addressing multimodal aspects of communication. It is compatible with the standard architecture, SAIBA (Situation Agent Intention Behaviour Animation) and integrates four different personality settings, which have been developed through the SEMAINE project: Spike (aggressive), Poppy (optimistic), Prudence (sensible) and Obadiah (miserable). It is capable of integrating gesture, facial expressions and prosody into communication. Although it has automatic generation of backchannels and gesture, like many systems it is lacking an automatic generation of prosody to supply to the TTS system

used (Mary Text To speech[1]) and could benefit from prosodic rules attached to dialogue act types to automatise this part of speech generation.

The paper is organised as follows: Section 2 presents the existing links between prosody, personality and dialogue acts that can be found in the literature, Section 3 shows the representation of dialogue acts in the SEMAINE corpus, Section 4 provides an analysis of the relation between speakers' personalities, their use of dialogue acts in this corpus and prosodic productions. Prosodic correlates of speech act types and the adequacy of speech act taxonomies are investigated in Section 5. Finally, we discuss the results in relation to the prosodic generation of dialogue acts for socio-affective ECAs.

2 Dialogue acts and prosody: related work

Dialogue act taxonomies and labelling. Dialogue management is a crucial part of animating a virtual agent. When engaged in interaction with a human user, the agent must be capable of interpreting the user's utterances, appropriately generating responses and successfully communicating intentions. The choice of taxonomy is very important and must be adapted to the task, and there must be clear distinctions between the different types of act. The first taxonomies for characterising communicative intentions were those developed in the context of Speech Act Theory by Austin (1962) and Searle (1979). Searle's classification of speech acts into five categories (assertives, directives, commissives, declarations and expressives) is still very influential today and is often used as the basis for communication models for virtual agents as well as for classification tasks. The taxonomic distinctions were based along three main dimensions: i) the purpose of the act, ii) the direction of fit between words and the world and iii) psychological states, although Searle also cites at least twelve criteria for distinguishing speech acts. The problem is that with so many dimensions, the task of annotating speech acts can be particularly difficult, especially if the criteria are subjective, such as assessing the psychological state of the speaker. For example, many expressives may also be considered assertions of an opinion and many assertives also spoken in an expressive manner, making the two acts particularly difficult to annotate. This apparent multifunctionality of speech acts and in particular the fact that it is often impossible to assign a single label is evoked by Allwood (1995). Due to their largely theoretical nature and their lack of cover of dialogue phenomena in real-life dialogue situations, speech act taxonomies were enriched in the form of dialogue acts, which cover a wider range of dialogue behaviour, including gesture, facial expression and non-linguistic verbal productions.

A large number of different taxonomies exist, with various structures from single-layered, mutually exclusive labels to multi-layered and hierarchical taxonomies (see (Popescu-Belis, 2003) for a comparison of differently structured taxonomies). Such a wide variety of taxonomies exist because labels are often domain-specific and the type of taxonomy is dependent on the goal of the task. For example, the dialogue acts used in the MapTask corpus (Anderson *et al.*, 1991) are specific moves based on the scenario of giving and following instructions to navigate a map. One taxonomy designed to be applicable across domains is the DAMSL taxonomy (Core and Allen, 1997), a multidimensional taxonomy designed for flexible and expressive annotation. However multidimensionality can be problematic for automatic approaches due to

[1] http://mary.dfki.de/

the sparsity of the combinations of dialogue act labels and also for annotation due to overlapping labels. DIT++ (Bunt, 2000), which is used as the basis for the norm ISO 24617-2 (Bunt *et al.*, 2012) aims to counter this problem by providing a hierarchically structured taxonomy from coarse to fine-grained categories. It is designed to be based on empirical distinctions and to provide easy decision-making when it comes to annotating dialogue acts. It provides the possibility of assigning multiple functions to individual functional units and of annotating a variety of multimodal phenomena. The use of a hierarchical structure means that the degree of precision can be adapted to particular tasks and coarser-grained decisions made if necessary. Very importantly, the taxonomy was designed to favour clearly defined clusters of acts that are mutually exclusive, which aids annotation decisions, and include only criteria that are empirically observable in dialogue.

Dialogue acts and prosody. The prosodic correlates of dialogue acts are of course dependent on the choices made in the taxonomy. However most taxonomies share some notion of statements, questioning and commanding. Traditionnally, researchers have studied the relationship between these acts and the sentence types declarative, interrogative and imperative. In the discussion on intonational contours associated with different speech acts in (Hirschberg, 2004), Hirschberg describes the standard contours for declaratives and wh-questions as being H* L-L% (i.e. with falling final pitch) and polar questions as being L* H-H% (i.e. with rising final pitch). However this one-to-one mapping of sentence type to speech act is rarely so clear-cut. For example, declarative questions are questions that use the declarative sentence type, and in spontaneous speech, context and not just sentence type is often needed to determine which act is intended. For example, Beun (2000) studies the use of context in determining dialogue act type. What is more, as discussed just before, dialogue act taxonomies are often more complex and contain labels that are better adapted to different dialogue situations.

The majority of studies identifying the prosodic correlates of dialogue focus on the improvement of automatic dialogue act classification and not on the study of prosody for dialogue generation, as is our aim here. Nevertheless a brief review of classification systems may be useful for identifying prosodic cues of dialogue acts, although the studies cannot necessarily provide guidelines as to how these features can be used for generating natural prosody. Shriberg *et al.*'s (1998) dialogue act classification model based on automatically extracted prosodic features related to pitch, duration, energy, pauses and speech rate was found to improve in accuracy, from 58.77% to 60.12%, with the addition of prosodic features, over a purely lexical model. Although this appears to be only a slight improvement, the difference was highly significant ($p < 0.001$), as verified by a Sign test. A more recent experiment by Hoque *et al.* (2007) also found that, whilst dialogue act classification based on prosody alone does not result in very high accuracy, prosodic features could be used to improve scores of a classifier based on discourse features, reaching an accuracy of 65.6% for a total of 13 speech act categories. Shriberg *et al.* use a decision tree, in which more salient features are likely to be placed higher up than less salient features. They find that utterance duration is particularly important and is used for decision-making in 55.4% of cases, most likely because of the correlation between duration and the number of tokens in the dialogue act. The next most salient feature is F0 at 12.6%, followed by pause duration (12.1%), energy (10.4%) and speech rate (9.4%). They also note that F0 gradient is particularly discriminating between questions and statements and that F0 and duration were the

two most salient features in the distinction between yes-no questions, wh-questions, declarative questions and statements.

Classification tasks provide invaluable information on the most salient prosodic cues for different dialogue acts. However the approach is not quite the same as those geared towards speech production, and the most salient features for classification are not necessarily the most perceptually salient features in terms of prosodic generation. Therefore studies in the identification of prosodic cues in view to generating speech are particularly important, especially since the generation of dialogue acts with appropriate prosody can afford to rely on more standardised and idealistic values, provided that these values are recognisable. One study that focuses on prosody and dialogue acts from this point of view is Syrdal and Kim's (2008) analysis of 12 hours of recorded speech by an American voice-over actress. According to their analysis, there were noticeable differences for average F0 values and pitch range across the various dialogue acts. The lowest average pitch was found for the act 'exclamation-negative', which also had the smallest pitch range (at 15Hz). The largest pitch range was found for requests at 163Hz and they also noticed general clusters of dialogue acts according to these two values. They observed a marked difference in speech rate for information-giving dialogue acts, depending on the detail of the information offered; a slower speech rate being adopted for more detailed information. No detailed quantitative analysis was performed of the general prosodic contours of the utterances, however the authors note several pertinent cases that go against syntactic intuition. For example the utterance "what was that?" would be traditionally classed under the syntactic unit 'wh-question', for which the prosodic contour is often noted in the literature as having a descending boundary tone, but in context and when representing a repetition, it takes on the prosodic contour usually associated with a yes-no question (with a low tone associated with "what" and an ascending boundary tone). Similarly, wh- and yes-no multiple choice questions were found to have similar prosodic contours, despite the fact that according to the traditional syntax-prosody mapping, they would be accorded very different intonations.

Personality and prosody. The relationship between personality and prosody is a growing field in voice synthesis, as expressive or emotional speech is often seen as a way of achieving more natural speech. This can be particularly useful in the field of virtual agents, where the aim could be to inspire confidence by presenting a friendly agent, or to place the user in a difficult situation by presenting a hostile one. One area that has received much attention is the creation of expressive corpora on which the speech synthesis can be based. For example, the SEMAINE corpus (McKeown *et al.*, 2012), to be used in our study, recreates a user-agent situation in which one of four roles is played by an operator. Each of the roles represents a different personality: Poppy (cheerful and outgoing), Prudence (pragmatic), Spike (aggressive) and Obadiah (pessimistic). These corpora can then be used as the basis for speech synthesis systems, without having to manually assign different prosodic settings to each personality. The disadvantage of this approach is the lack of control over the prosodic settings, especially if multiple factors (such as the choice of dialogue act) influence prosody. It can therefore be useful to look at the prosodic correlates of different personalities to be able to control this aspect independently of the initial stages of speech synthesis by applying post-processing prosodic rules. Several studies have been performed on identifying vocal cues corresponding either to discrete personality roles or to emotional dimensions. For example, Laukka *et al.* (2005) perform a detailed acoustic

analysis of five different emotions: 'anger', 'fear', 'disgust', 'happiness' and 'sadness' and they found that for the excitement dimension, high levels of excitement are often indicated by a high pitch level and range and higher intensity, but that vocal cues were often correlated with several different dimensions, indicating that personality is a complex mix of prosodic factors. Similarly, Scherer (2003) writes on the importance of the portrayal of emotion in speech and reviews the literature on the relationship between emotion and prosody. He too makes the link between sadness and low intensity, pitch levels and range and speech rate, compared to high levels for joy/elation, but remarks that recognition accuracy of emotions through voice alone is relatively low, especially compared to recognition through facial expressions, suggesting the complex nature of the interpretation of emotion through vocal cues.

3 Methods

We performed a corpus analysis with the ultimate task of speech synthesis in mind. The reference corpus used here was the SEMAINE corpus (McKeown *et al.*, 2012), an emotionally-coloured conversational database consisting of dialogues between an operator and a user. The chosen scenario was that of the Solid Sensitive Artificial Listener, in which dialogue between a person playing the role of the operator and a user was recorded visually and auditorily. Dialogue was non-scripted to allow for the most spontaneous dialogue and gestures possible, and so the lengths of the sessions are variable depending on the ease of conversation and the loquaciousness of the participants. The only constraint was that the operator was unable to reply to questions.

The operator took on one of four roles: Poppy (cheerful and outgoing), Prudence (pragmatic), Spike (aggressive) and Obadiah (pessimistic). The aim of SEMAINE was to produce a corpus of emotionally-charged dialogue, collected through the activity of conversation, with both verbal and non-verbal dialogue behaviour. For each session the user and operator were situated in separate rooms, equipped with video screens and recorded using wearable microphones, with audio recorded at 48 kHz and 24 bits per sample. In light of our task of providing guidelines for generating speech in an ECA, only the operator's speech was analysed.

The aim of the analysis was two-fold:

1. to identify whether there exist clearly identifiable prosodic correlates for each of the speech act types and to compare two differently structured taxonomies to test dialogue act distinctions in terms of prosodic correlates for dialogue generation.

2. to determine how personality types interact with these dialogue acts to influence the different stages of dialogue act generation.

A sub-corpus of SEMAINE was randomly selected, consisting of just under three hours of dialogue divided into 36 sessions. The sessions were chosen to ensure the same number of sessions per speaker and per role. Given the spontaneous nature of the dialogue, session times varied from two to eight minutes with an average duration of just under five minutes. However in practice, total durations were comparable between speakers and between the different roles. One female and two male researchers

from Belfast performed the role of the operator and each of the four personality types is represented by three sessions for each speaker.

3.1 Dialogue act annotation and segmentation

The sub-corpus was manually annotated for dialogue acts according to two different taxonomies. The first taxonomy used was Searle's taxonomy of five speech acts, although due to the nature of the dialogue, only assertives, directives and expressives feature in the corpus. Their distribution according to speaker can be seen in Table 1. The second taxonomy focuses on the division of directive acts into 12 different acts based on communicative intention and inspired by Bunt's DIT++ taxonomy (2000). Both sets of annotations were produced by the first author, a native speaker of English, as a preliminary study in view to launching a full-scale annotation campaign.

	Assertive	Directive	Expressive	Total
F1	79	180	156	415
M1	158	111	99	368
M2	107	124	113	344
Total	344	415	368	1127

Table 1: The distribution of the three different types of dialogue act among the three speakers (F1, M1, M2) according to Searle's taxonomy of speech acts

The size and scope of the units considered for dialogue acts are often disputed, and here a simplified notion of functional segment was used for both taxonomies to facilitate the prosodic analysis and to focus uniquely on certain aspects of speech. Whereas in DIT++, functional units can be discontinuous and overlapping if a single utterance is multifunctional (Bunt, 2011), such an analysis is poorly adapted to analysing prosodic features of utterances and especially for studying pitch and intensity contours for which the sequential aspect requires the analysis of a continuous segment.

Incomplete utterances and speech disfluencies that interrupted the prosodic flow of the utterance were not included for annotation, based on the fact that they disrupt the continuity of prosodic production. Although they represent a normal aspect of human speech and would be important for classification tasks, they may be considered less pertinent for our task of generating dialogue acts, where the aim is to identify characteristic prosodic parameters, and would therefore constitute part of a separate study. The filtering did not affect manual dialogue act annotation since the annotator had access to the complete dialogue.

This also ensures that the segments annotated are as close as possible to the expected input to the TTS system, containing for the most part well-formed sentences. The nature of the SEMAINE corpus, in which actors fulfilled specific personality-based roles based on a simple agenda, meant that utterances contained fewer of these incomplete or highly disfluent utterances than would have been expected in completely spontaneous conversations. We estimate the percentage of utterances not to have been annotated for this reason to be approximately 3%. The segments taken were segments often corresponding to a sentential unit and where possible to intonational phrases. Tag questions were segmented into two separate parts, the first consisting

of a statement-like sentence, and the second of the tag containing an auxiliary and a pronoun, which was annotated as a question.

3.2 Functional labels for the second taxonomy

A second taxonomy was used in order to test the hypothesis that further taxonomic distinctions may be necessary to account for certain important prosodic differences, especially within the category of directives. An analysis of the different types of utterances grouped under the term 'directives' identifies the types shown in Table 2 (defined in terms of intention and inspired in part by Harry Bunt's DIT++ taxonomy).

Dialogue act	Number	Comments
Infoseek wh	118	Wh- questions
Infoseek yn invert	125	Yes-no questions with subject-verb inversion
Infoseek yn noinvert	34	Yes-no questions without subject-verb inversion
Infoseek yn part	25	Yes-no questions formed of an incomplete sentence (no subject or verb)
Infoseek choice	1	Choice question (with 'or')
Tag	3	Tag question undefined for polarity
Tag positive	3	Tag question in which the speaker believes the preceding act to be true
Tag negative	9	Tag question in which the speaker is questioning the veracity of the previous act
Infoseek command	31	Requesting information through a command ('Tell me…')
Infoseek assertive	8	Response-seeking assertion ('I heard that you have moved recently', 'I might feel better if you told me')
Advice	24	Advising the addressee about what is best ('You should…', 'I advise you…', 'My advice is…', 'What I would do is…' etc.)
Suggestion	12	Suggesting an activity or action or inviting the user to do something ('Let's do…', 'How about if we do…' etc.)
Command	22	Commands other than information-seeking commands

Table 2: The distribution of directive acts into sub-types (in number of acts)

This list is not meant to be exhaustive, but a demonstration of the very different forms of utterance that can be grouped under a single label. Whilst assertives and expressives generally have the form of declarative sentences, directives can regroup declaratives, interrogatives and imperatives.

3.3 Prosodic features

The time-coded transcription was used to generate textgrids using the Praat Software (Boersma and Weenink, 2014), where were verified manually. A range of prosodic features were extracted, including pitch and intensity values (mean, max, min, range,

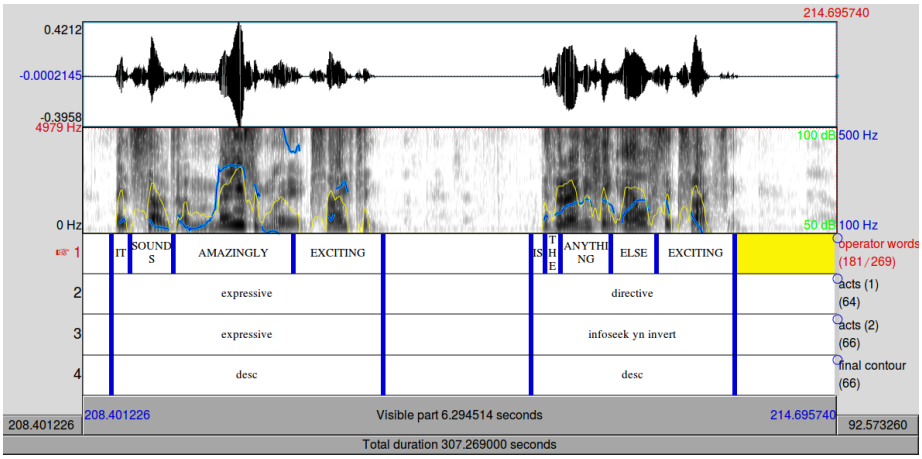

Figure 1: An audio session with four tiers of aligned annotations. The second and third tiers correspond to the first and second taxonomies respectively

standard deviation), number of pauses, average length of pauses, speech rate (words per second) and average pitch slopes[2].

To complete these general values, all dialogue acts were manually annotated for the final pitch contour, based on the pitch changes over the final word of each utterance. When describing intonational contours, there are several models that are commonly used. One option is to automatically extract F0 values that make up the contour. Although this approach is fast and easy, it has the disadvantage of being more difficult to interpret because of the micro-distinctions made, the lack of stylisation and also the inexactness of F0 extraction, especially at the end of utterances due to the use of creaky voice, which is often characteristic of falling final tones. Another option is to describe pitch contours in terms of stylised descriptors such as fall, rise-fall or fall [for example the methods of the British School, based on the work of Palmer (1922)]. Unlike the F0 values, this method has the advantage of using relative descriptors rather than absolute values, which could be more appropriate when comparing different speakers. A very widely used formalism for contour-coding is the ToBI system (Tones and Break Indices) described in (Silverman *et al.*, 1992) and based on Pierrehumbert's theoretical approach (Pierrehumbert, 1980). The system is more elaborate and uses phrase tones (initial and final boundary tones), pitch accents and break indices to indicate prosodic changes. For example, as mentioned previously, H* L-L% corresponds to a high pitch accent followed by a low boundary tone, indicating a falling final pitch contour. The disadvantage of such an approach is that annotation is time-consuming and inter-annotator scores unreliable for spontaneous speech; Yoon *et al.* (2004) perform an inter-annotator analysis of ToBI annotations on spontaneous telephone conversations and a kappa coefficient of 0.48 is obtained for the choice of phrasal accent, although a more adequate coefficient of 0.79 is obtained for the choice of boundary tone. We therefore choose to perform a simple contour analysis, similar to the ToBI boundary tone analysis but using the stylised descrip-

[2]Note that certain of these values, notably the pitch and intensity values, are sensitive to recording conditions and so comparison with previous works should be made with care. Here we compare only the values from the SEMAINE corpus between each other.

tors ‘rise’, ‘fall’, ‘rise-fall’, ‘fall-rise’, ‘flat’ and ‘other’. These descriptors were based on perceptual judgments although the annotator could also refer to the visual pitch representation on the textgrid, with full knowledge that the detection of pitch contours via Praat is prone to errors. An example textgrid with all annotations can be seen in Figure 1.

4 Results: influence of personality on dialogue act choice and on prosodic production

Although choices of act are also speaker-dependent, which is why the results below are separated by speaker, the number and percentage of different acts differ according to which character is being played, suggesting that personality does have an influence on dialogue behaviour.

Personality	Assertive	Directive	Expressive	Total
F1 Obadiah	34 (36%)	26 (28%)	34 (36%)	94
F1 Poppy	13 (11%)	62 (54%)	40 (35%)	115
F1 Prudence	17 (15%)	49 (42%)	50 (43%)	116
F1 Spike	15 (17%)	43 (48%)	32 (36%)	90
M1 Obadiah	76 (67%)	11 (10%)	26 (23%)	113
M1 Poppy	31 (28%)	48 (43%)	33 (29%)	112
M1 Prudence	25 (34%)	29 (39%)	20 (27%)	74
M1 Spike	26 (38%)	23 (33%)	20 (29%)	69
M2 Obadiah	28 (50%)	17 (30%)	11 (20%)	56
M2 Poppy	22 (23%)	41 (44%)	31 (33%)	94
M2 Prudence	19 (25%)	21 (28%)	36 (47%)	76
M2 Spike	38 (32%)	45 (38%)	35 (30%)	118

Table 3: The number of each act type (and percentages of total acts) for each speaker-personality combination. Totals may not equal 100% due to rounding.

The classification according to Searle’s taxonomy (shown in Table 3) shows that for all three speakers Obadiah produced the highest percentage of assertive acts and is also associated with amongst the lowest percentage of directive acts out of the three personalities. Prudence was associated with the highest percentage of expressive acts for two out of the three speakers and Poppy with a higher percentage of directives for all three speakers. These differences could be seen as the consequence of the different personalities of the operator, Obadiah’s relatively few directive acts being linked to a lack of engagement with the addresse and the fact that Poppy is associated with a higher production of directives could be related to a higher level of engagement with the user.

However what is more interesting is the choice of act amongst the different types of directive act, as annotated in the second taxonomy and as shown in Table 4. Certain acts such as advice or suggestions are too rare to be analysed. However certain different strategies can be seen, notably in the different ways of asking a question. For all three speakers, Obadiah does not produce a single yes-no part production (corresponding to an averbal yes-no production) and the majority of the yes-no questions produced as Obadiah are yes-no questions with inversion. Along with the fact that

	Infoseek wh	Infoseek yn invert	Infoseek yn noinvert	Infoseek yn part	Infoseek choice	Tag undefined	Tag positive	Tag negative	Infoseek command	Infoseek assertive	Advice	Suggestion	Command
F1 Obadiah	10.6	6.4	3.2	0	0	1.1	2.1	0	0	1.1	1.1	2.1	0
F1 Poppy	13.9	21.7	3.5	4.3	0	0	0.9	0.9	4.3	0.9	0	3.5	0
F1 Prudence	12.1	19.0	3.4	2.6	0	0	0	0	2.6	0.9	1.7	0	0
F1 Spike	18.9	7.8	8.9	6.7	0	0	3.3	0	1.1	1.1	0	0	0
M1 Obadiah	0	2.7	0	0	0	0	0.9	0	2.7	1.8	1.8	0	1.8
M1 Poppy	10.7	14.3	1.8	1.8	0	0	0.9	0.9	5.4	0.9	0	1.8	4.5
M1 Prudence	5.4	17.6	0	1.4	0	1.4	0	0	2.7	0	6.8	1.8	2.7
M1 Spike	4.3	2.9	4.3	1.4	1.4	0	0	1.4	4.3	0	0	1.4	11.6
M2 Obadiah	14.3	5.4	0	0	0	0	0	0	1.8	1.8	5.4	0	1.8
M2 Poppy	8.5	14.9	6.4	2.1	0	0	0	0	2.1	0	8.5	0	1.1
M2 Prudence	10.5	9.2	0	2.6	0	0	0	0	0	0	3.9	1.3	0
M2 Spike	15.3	5.9	3.4	2.5	0	0.8	0.8	0	4.2	0	1.7	0.8	2.5

Table 4: The distribution of different types of directive act for each speaker-personality combination (in percentages of the total dialogue acts annotated)

Obadiah is associated with few directive acts in general, this could be linked to a more formal style in which the operator questions less what the user is saying, especially since a number of the yes-no part productions produced by other roles are repetitions of the previous speech turn. Spike on the other hand was the only personality associated with more yes-no questions without inversion than with (for all three speakers). Given that his is an aggressive personality, a high percentage of yes-no non-inverted questions could be associated with a higher level of directness and even a less polite or more familiar style of speech, which provides interesting leads for further research.

Based on the prosodic features that have been automatically extracted with Praat [see Section 3.3), an analysis of the prosodic correlates of personality has been carried out (See Figure 2]. It reveals a couple of interesting and consistent points concerning the relationship between personality and prosody:

- Spike (aggressive personality) was the personality associated with the highest intensity values.
- Obadiah (pessimistic personality) was most often associated with the slowest speech rate of the four personalities.
- Poppy (cheerful personality) and Prudence (pragmatic personality) were expressed using the greatest pitch variation and range of the four personalities, and Poppy with the highest average pitches.

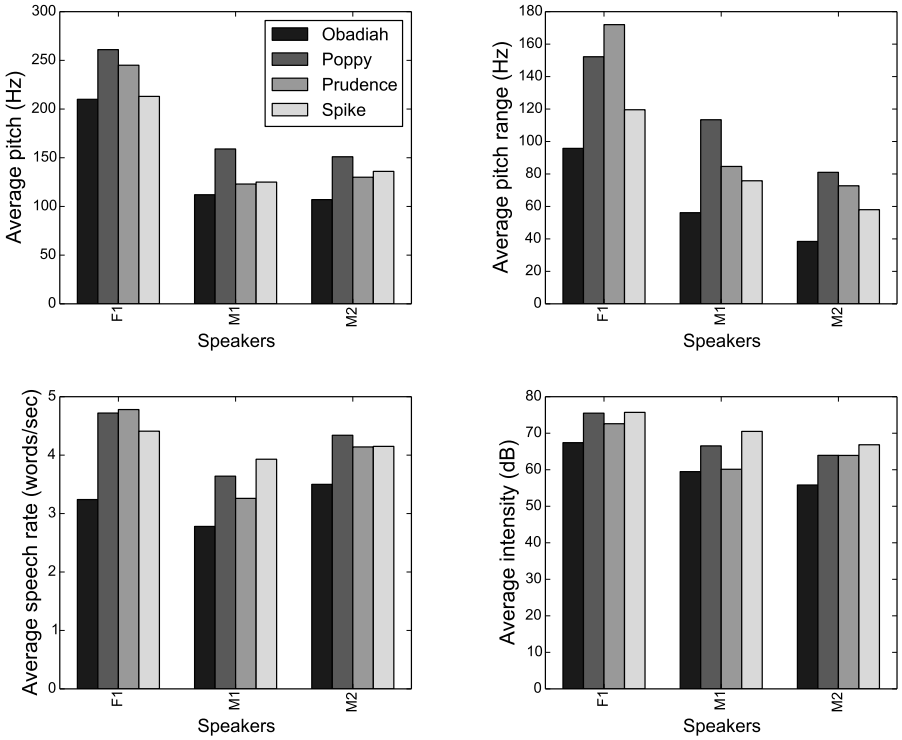

Figure 2: Prosodic correlates of personality, by speaker (average pitch, average pitch range, average speech rate and average intensity)

5 Results: intonation of dialogue acts

The analysis of the prosodic correlates of the different dialogue acts reveal that the choice of taxonomy is particularly important when distinguishing between different types of act.

Speech act	Rise	Fall	Rise-fall	Fall-rise	Flat	Other
Assertive	7	84	5	2	2	0
Directive	31	58	4	3	3	0
Expressive	8	80	3	2	6	1

Table 5: The percentages of speech acts identified for each of the contour types. Totals may not equal 100% due to rounding

The analysis based on Searle's taxonomy (see Table 5) showed that very few prosodic differences were found between assertive and expressive acts. For all three speakers, expressives were characterised by a slightly higher average pitch and for two speakers a higher pitch range than assertives. However no differences were found in terms of speech rate, intensity or pitch slope. Moreover, the manual annotations of final contours suggest very similar prosodic contours for these two types of act,

the most predominant contour being descending, at 84% for assertives and 80% for expressives[3]. This could suggest that the main difference between these two acts is essentially in terms of other factors such as the lexical content, expressives containing more emotionally-charged words than assertives. Directives were found to be expressed by all three speakers with a higher intensity than the other two types of act and with more pitch variation. This could correspond to the fact that these acts are intended to provoke an addressee response and need to draw the addressee's attention. Although the average pitch slope of directives was descending, as with the other two types of act, the annotations of final pitch contours suggest a far more heterogeneous type of act than the other two.

Speech act	Rise	Fall	Rise-fall	Fall-rise	Flat	Other
Infoseek wh	12	83	0	1	3	1
Infoseek yn invert	40	34	7	4	1	0
Infoseek yn noinvert	50	35	0	6	9	0
Infoseek yn part	80	12	8	0	0	0
Infoseek choice	0	100	0	0	0	0
Tag	100	0	0	0	0	0
Tag positive	11	78	0	0	11	0
Tag negative	100	0	0	0	0	0
Infoseek command	3	87	3	3	3	0
Infoseek assertive	0	87	0	0	13	0
Advice	8	75	8	4	0	4
Suggestion	0	75	25	0	0	0
Command	0	82	5	5	9	0

Table 6: The percentages of speech acts identified for each of the contour types. Totals may not equal 100% due to rounding

The re-annotation of directive acts according to the second taxonomy resulted in the separation of the different intentions (notably questioning, commands, advice) associated with the rather heterogenous act 'directive', and these distinctions, especially concerning the questions, proved useful in the analysis of final contours (see Table 6), since more distinct generalisations occur in the final pitch contours of certain acts. Certain observations, such as a majority of descending contours for wh-questions, confirm generalisations cited in the literature. However although yes-no questions are often said to be globally rising, a very high percentage have final descending contours. Yes-no questions formed of an averbal phrase (infoseek yn part), often for confirmations of the previous turn, are indeed classed as having a majority of final rising contours. However 34% of yes-no questions with subject-verb inversion have descending contours, which must be accounted for in generating prosody. Yes-no questions without subject-verb inversion (for which a final rising contour is often said to be the factor marking them as questions as opposed to statements) also have a large percentage of descending final contours (35%). On closer inspection, lexical factors appeared to play a role in the successful communication of these utterances as

[3]Although no comparison is made in this study with other accents, the Belfast accent is known to be associated with high rising terminal inflections in declarative sentences, which could mean that a higher percentage of assertives and expressives are associated with a rising intonation than for example with SSE accents.

questions. For example, the majority of these acts were introduced by the word 'so' or concerned the addressee and his/her desires, beliefs or actions, for example "you just let them walk all over you" or "so you're taking it box by box"[4]. A notable difference was also found between negative and positive tag questions, the former being characterised by a rising intonation and the latter by a descending intonation. This distinction appears to be based on the degree of certainty, as evoked by Gravano *et al.* (2008) for the correlation between certainty and downstepped pitches and Ŝafárová (2006) for the correlation between uncertainty and rising final pitch.

6 Discussion

The analysis of the prosodic correlates of dialogue acts showed the importance of finer-grained distinctions in the category of directives and also of the lack of distinctions between expressive and assertives acts, which was also reflected in the difficulty in annotating these two acts, as mentioned in Section 2. Certain observations, such as a majority of descending contours for wh-questions confirmed those made in the literature, however certain observations deviated from these generalisations. For example, a high percentage of yes-no questions were classed as having a descending final contour, despite the fact that they are said to be produced with ascending contours. It is particularly interesting that the yes-no questions with no characteristic subject-verb inversion were not necessarily produced with a final ascending contour, despite the fact that this prosodic contour is often said to be the factor that marks these acts as questions rather than assertives. A secondary preliminary analysis based on the subcorpus used suggests that lexical factors such as the introductory word 'so' or the subject content concerning the addressee can be used in these cases to indicate that these acts are indeed reponse-provoking acts.

More generally, the present study provides some research ideas for the generation of dialogue acts for socio-affective ECAs. In terms of naturalness, it is important to establish generalised correspondences between dialogue acts, personality and prosody, but also to enable variation within these types, the aim being to successfully convey intention and emotion in the most natural way possible. Even though more results would be needed to make across-the-board generalisations and to apply appropriate statistical tests to corroborate our analyses, personality was seen in this study to have a certain effect on the choice of dialogue acts and their prosodic production. It was not possible in this study to compare the simultanenous effects of personality and dialogue act on prosody due to the sparsity of the data, but this will be the aim of future works.

The first issue is now how to generate the relevant prosody according to dialogue acts and socio-emotional factors such as the ECA's personality based on the previous analyses. Existing speech synthesis systems, such as Mary Text To speech, use the ToBI convention to model prosody in their system. Although certain TTS systems do allow more integration of emotion and expressivity, the link between dialogue acts,

[4]This phenomenon has previously been noted in a corpus study of Dutch dialogues by Beun (1989), where only 48% of declarative sentence had rising intonation, and there was seen to be a correlation between the use of the second person personal pronoun and of particles such as 'and' and 'so' in the identification of questions from declarative sentences with falling intonation. See (Šafárová, 2006) for an example of the use of the 'you'-pronoun and particles to indicate response-seeking acts in American English. A further semantic explanation for the lack of a rising intonation in declarative questions is provided by Beun (2000), which suggests that a greater degree of certainty in the demand for confirmation is linked with a greater probability of a descending final contour.

prosodic parametrisation and such expressivity is rarely described. The corpus-based prosodic analysis provides a basis for the specification of prosodic rules according to the different dialogue acts used to portray different intentions and emotions.

Even more important for the generation of dialogue acts for socio-affective ECAs is the integration of prosodic choice in a full sequential dialogue model. It leads us to the second issue, which is how to gain full control over prosodic rules. This involves integrating them into the dialogue context and synchronising the prosodic choice with the ECA's socio-affective model and other modalities such as gesture. Analysing face-to-face interactions in a human-agent scenario is one of the solutions to this issue.

Finally, dialogue act generation should also be studied in relation to Natural Language Generation (NLG). The semantic content of the utterance as well as any features indicating the way in which it is to be expressed are used in NLG to provide the most appropriate expression given the intention and other external factors such as personality and emotion. A particularly challenging task in NLG is to provide a range of different expressions for a single logical form, allowing for naturalness in the form of variation of structure and the portrayal of certain communicative behaviours such as familiarity, hostility, openness, etc. Our corpus analysis could therefore be extended to study the link between semantics and prosody.

7 Conclusion

Having analysed the emotional corpus SEMAINE, it has become clear that the decisions made in the choice of a taxonomy are important for making prosodic distinctions between dialogue acts. Not only do the distinctions made need to be easily identifiable and sufficient in terms of generating different types of intention, they also need to encode a sufficient amount of information to make useful distinctions later in the system, notably in terms of prosodic generation. The taxonomy must also be adapted to the task in hand and be able to model a multitude of different communicative intentions that are also appropriate for multimodal behaviour such as speech, gesture, facial and head movements. The use of a richer taxonomy such as one offered by a subset of DIT++, which enables this modality and clearly separates functions by dimension and by intention, is a necessary addition to any dialogue act system. Using this more detailed dialogue act taxonomy, we have also succeeded in making a first step towards defining differences in both the choice of dialogue act and in the prosodic realisation of these acts based on an ECA's personality.

References

Allwood, J. (1995). An activity based approach to pragmatics. Technical report, Department of Linguistics, University of Göteborg.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, **34**, 351–366.

Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, H., Jones, H., Ochs, M., C., P., Porayska-Pomsta, K., Rizzo, P., and Sabouret, N. (2013). The TARDIS framework: Intelligent virtual

agents for social coaching in job interviews. In *Proceedings of the 10th International Conference on Advances in Computer Entertainment*, Heidelberg. Springer.

Austin, J. L. (1962). *How to Do Things With Words*. Clarendon Press, Oxford.

Beun, R.-J. (1989). Declarative question acts: Two experiments on identification. In M. M. Taylor, F. Néel, and D. Bouwhuis, editors, *The structure of multimodal dialogues*. North Holland, Amsterdam.

Beun, R.-J. (2000). Context and form: Declarative or interrogative, that is the question. In H. Bunt and W. Black, editors, *Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics*. John Benjamins, Amsterdam.

Bevacqua, E., Prépin, K., Niewiadomski, R., De Sevin, E., and Pelachaud, C. (2010). Greta: Towards an interactive conversational virtual companion. In Y. Wilks, editor, *Artificial Companions in Society: perspectives on the Present and Future*, pages 143–156. John Benjamins, Amsterdam.

Boersma, P. and Weenink, D. (2014). Praat: doing phonetics by computer (version 5.3.51) [computer program]. retrieved 31 May 2014 from http://www.praat.org.

Bunt, H. (2000). Dynamic interpretation and dialogue theory. In M. M. Taylor, D. Bouwhuis, and F. Néel, editors, *The Structure of Multimodal Dialogue*, volume 2, pages 139–166. John Benjamins, Amsterdam.

Bunt, H. (2011). Multifunctionality in dialogue. *Computer Speech and Language*, **25**(2), 222–245.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, C. A., Hasida, K., Petukhova, V., Popescu-Belis, A., and Traum, D. (2012). ISO 24617-2: Semantically-based standard for dialogue annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Campano, S., Clavel, C., and Pelachaud, C. (2015). I like this painting too: When an ECA shares appreciations to engage users. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1649–1650.

Core, M. G. and Allen, J. F. (1997). Coding dialogs with the DAMSL annotation scheme.

De Carolis, B., Pelachaud, C., Poggi, I., and Steedman, M. (2004). APML, a mark-up language for believable behavior generation. In H. Prendinger and M. Ishizuka, editors, *Life-like Characters. Tools, Affective Functions and Applications*, pages 65–85. Springer, Heidelberg.

Gravano, A., Benus, S., Hirschberg, J., Elisa, S. G., and 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.

Hirschberg, J. (2004). Pragmatics and intonatinon. In L. R. Horn and G. Ward, editors, *The Handbook of Pragmatics*. Blackwell, Oxford.

Hoque, M. E., Sorower, M. S., Yeasin, M., and Louwerse, M. M. (2007). What speech tells us about discourse: The role of prosodic and discourse features in speech act classification. In *IEEE International Joint Conference on Neural Networks*, pages 2999–3004, Orlando, FL.

Klatt, J., Marsella, S., and Krämer, N. (2011). Negotiations in the context of AIDS prevention: An agent-based model using theory of mind. In H. Vilhjálmsson, S. Kopp, S. Marsella, and K. Thórisson, editors, *Intelligent Virtual Agents*, volume 6895 of *Lecture Notes in Computer Science*, pages 209–215. Springer, Heidelberg.

Laukka, P., Juslin, P., and Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, **19**(5), 633–653.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schröder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, **3**(1), 5–17.

Palmer, H. E. (1922). *English Intonation with Systematic Exercises.* Cambridge, Heffer.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, MIT.

Popescu-Belis, A. (2003). Dialogue act tagsets for meeting understanding: An abstraction based on the DAMSL, Switchboard and ICSI-MR tagsets.

Ŝafárová, M. (2006). *Rises and falls: Studies in the semantics and pragmatics of intonation.* Ph.D. thesis, Institute for Logic, Language and Computation. 59-74.

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication*, **40**(1-2), 227–256.

Searle, J. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts.* Cambridge University Press.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., and Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech (Special Issue on Prosody and Conversation)*, **41**(3-4), 439–487.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In J. Ohala, T. Nearey, B. Derwing, H. M.M., and G. Wiebe, editors, *Proceedings of the International Conference on Spoken Language Processing*, pages 867–870, Department of Linguistics, University of Alberta.

Suignard, P. (2010). NaviQuest: un outil pour naviguer dans une base de questions posées à un agent conversationnel. In *Workshop sur les Agents Conversationnels Animés*, Lille, France.

Syrdal, A. and Kim, Y.-j. (2008). Dialog speech acts and prosody: Considerations for TTS. In *Proceedings of Speech Prosody*, pages 661–665, Campinas, Brazil.

Yoon, T.-j., Chavarría, R., Cole, J., and Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2729–2732, Nara, Japan.