SYNTHESISING CONTEXTUALLY APPROPRIATE INTONATION IN LIMITED DOMAINS

Rachel Baker, Robert A. J. Clark and Michael White

CSTR/HCRC, The University of Edinburgh

ABSTRACT

We describe a method of synthesising contextually appropriate intonation with limited domain unit selection voices. The method enables the natural language generation component of a dialogue system to specify its intonation choices via APML, an XML-based markup language. In a pilot study, we built an APML-aware limited domain voice for use in flight information dialogues, and carried out a perception experiment comparing the APML voice to a default version built using the same recordings without the additional structure. The intonation produced by the APML voice was judged significantly more contextually appropriate than that of the default voice. These results justified building a second voice with a much larger vocabulary, using an automated script generation algorithm.

1. INTRODUCTION

Unit selection, the concatenation of larger than segmental units of speech from a database using sophisticated searching and joining algorithms, is the method of choice for high-quality speech synthesis [1]. However, while unit selection synthesis typically offers high quality, it does not allow much control over the intonation contour of an utterance. In this paper, we investigate a method of adapting current limited domain synthesis techniques [2] in order to produce intonation that is more contextually appropriate. The motivation for our study comes from the speech output needs of the FLIGHTS dialogue system [3]. In the FLIGHTS system, the natural language generator consults decision-theoretic user models and the dialogue history in order to generate tailored, context-sensitive descriptions of the available options. These descriptions compare and contrast the most compelling attributes of the most relevant flights, rather than simply listing query results. In a system that is able to help users sort through the options in this way, we hypothesise that it becomes especially important to enable the generator to specify intonation that expresses contrast intelligibly.

1.1. Information Structure and Intonation

The user preferences in FLIGHTS affect many aspects of how the output content is selected, organised and expressed. For example, if a user cares most about price, the system may refer to a flight as *the* CHEAPEST *flight* (if the flight in question is indeed the least expensive of the relevant options). As another example, for a user that prefers to fly business class, the system may identify interesting trade offs with previously mentioned flights using both prosodic emphasis and discourse cues such as *but*, as in *There* ARE *seats in business class on the British Airways flight that arrives at four twenty p.m., but you'd need to connect in Manchester*.

The FLIGHTS generator employs Steedman's [4, 5] theory of information structure in choosing contextually appropriate intonation. Steedman's theory is based on two primary distinctions: theme vs. rheme, and marked vs. unmarked. A theme is "what the participants have agreed to talk about" [6, p.2], the part of the sentence that ties it to the previous discourse; a rheme is the speaker's new contribution on the subject of the theme. Marked information is either new or contrastive; unmarked information is neither.[1]

Each theme and rheme is potentially a member of a theme- or rheme- *alternative set* (TAS and RAS). An alternative set is made up of all the phrases (or more precisely, the semantics thereof) that could have appeared in the same position. For example, in the answer of the question-answer pair:

Q: When does the CHEAPEST flight leave?
A: (The CHEAPEST flight leaves) (at SEVEN A.M.)

The rheme, *at* SEVEN A.M., could have been replaced by *at* EIGHT A.M. or *at* SEVEN P.M., so these are all members of the RAS. The parts of the rheme that help to distinguish it from other members of the RAS—here, the hour and a.m./p.m.—are marked, while the rest of the rheme is unmarked. Likewise, the parts of the theme that distinguish it from other members of the TAS are marked—here, the property of the flight being cheapest distinguishes this theme from those involving other contextually salient flights (e.g., the DIRECT one).[2] It is not uncommon to have single-

[1] In [4] and earlier work, Steedman uses the term *focus* (in the narrow sense) for markedness; to help avoid terminological confusion, in [5] he switches to the term *kontrast*, following Vallduví and Vilkuna [7]. Here we simply use the term 'marked', less technically.

[2] Indeed, the need for theme accents becomes clearer when such contextually available alternatives appear explicitly, e.g. in *I know when the* DIRECT *flight leaves, but when does the* CHEAPEST *flight leave?*

ton TASs, containing only the actual theme; these themes are entirely unmarked.

Steedman's theory assigns different kinds of pitch accents to marked words based on whether they are part of a theme or a rheme. Marked words in themes generally receive an L+H* pitch accent, but it is also possible for them to receive an L*+H accent [5]. Marked words in rhemes generally receive an H* pitch accent, but can also receive L*, and possibly H*+L, and H+L*. H* and L+H* are used in rhemes and themes that are denoted by the speaker as "agreed", or uncontentious. L* and L*+H, in contrast, are used in themes and rhemes which the speaker claims are contentious, because either the speaker or the hearer is not committed to them [5]. L* tends to be used in the quite specific situations in which a negative answer is expected or when politely listing alternatives [4]. Because of the generally cooperative nature of the flight information domain, we have limited our attention to only the more general H* pitch accent for marked words in rhemes, and only L+H* for marked words in themes.

Phrase accents appear at the end of intermediate phrases and boundary tones appear at the end of intonational phrases, which consist of one or more intermediate phrases. Steedman's explanation of the relation between phrase accent or boundary tone type and information structure has changed somewhat over the years. In [5], he ties boundary tones in with turn-taking. H, LH% and HH% mark hearer responsibility for, or commitment to, the information unit. L, LL% and HL% mark speaker commitment to the unit. Commitment to a unit can involve previous knowledge of or belief in the information in the unit, or the speaker's belief that the hearer will believe the information having heard it. As with the question of agreement, speaker and hearer commitment is based entirely on the speaker's claims about each of their commitments, rather than the actual set of beliefs they possess. Once again, we have employed a simplified version of the theory because of the restricted nature of the sentence types in the FLIGHTS domain. It builds on [4], which associates the contour H* L with rhemes and L+H* LH% with themes. When a rheme comes at the end of a sentence, and therefore at the end of an intonational phrase, it is marked with H* LL%. When a rheme ends with a comma, it is sometimes marked with H* LH% to communicate the speaker's intention to continue speaking in spite of the pause that often accompanies commas.

Steedman's information structural semantics of intonation is fully integrated into Combinatory Categorial Grammar (CCG). This grammar integrates intonation structure into surface derivational structure, notably even when the intonation structure departs from the restrictions of traditional surface structure. In the FLIGHTS generator, the theme/rheme and markedness choices made by the content planner determine the intonation choices made by the Open-

```
<apml>
   <performative type="inform">
   <rheme>
     The <emphasis x-pitchaccent="Hstar">KLM</emphasis>
     Airlines flight <boundary type="L"/>
   </rheme>
   <theme>
     leaves <emphasis x-pitchaccent="LplusHstar">
     Edinburgh</emphasis> at
     <emphasis x-pitchaccent="LplusHstar">eleven
     a.m.</emphasis> <boundary type="LH"/>
    </theme>
  </performative>
</apml>
```

Fig. 1. APML for an example sentence

CCG realiser [8, 9], which contains a suitable implementation of CCG for English. The realiser's intonation choices are conveyed to Festival for synthesis via APML [10], an XML-based markup language for specifying turn-taking, performative, affective, and intonational aspects of text. An example extract is shown in Fig. 1.

1.2. Using APML in Limited Domain Synthesis

Festival's cluster unit synthesis uses techniques based on [1]. Units are clustered into groups of acoustically similar units based on non-acoustic information available at synthesis time. A pre-built classification and regression tree [11] replaces the target cost, thus reducing computational load.

The actual choice of unit type is one of the most important factors with cluster unit synthesis, the default choice often being phone name in combination with the word that the phone is from. This means that a 't' from the word *table* would not be used to synthesise the word *tablet*. The disadvantage of this is that the synthesiser can only speak words that exist in the original data set. This restriction is not a problem when the input is a known vocabulary originating from a natural language generation component. Additionally, the restriction greatly improves synthesis quality because co-articulatory effects and allophonic variations specific to the pronunciation of a certain word are automatically preserved.

This project takes the default unit type restriction of phone from a given word one step further, so not only must a phone come from the word to be synthesised, it must also come from a word with the same predicted pitch accent and boundary tone type. The revised unit type takes the form phone_word_pitch-accent_boundary-tone. The pitch accent category can take the value H*, L+H*, or NONE; the boundary tone category can be LH, LL, L, or NONE.

This approach differs from other work in that although the pitch accents and boundaries are ToBI labels, they do not necessarily reflect the actual shapes of the intonation contours for the recorded words. The labels are theoretical, idealised representations of the intonation predicted by

the information structural status of the word. In building the voice, we have used the labels predicted by the theory for each recorded prompt, rather than labelling the prompts following ToBI guidelines. The idea is that even if the F0 contour of a focused word does not take the exact form predicted by the theme/rheme theory, its status will be marked prosodically, and this marking can be consistently carried over into synthesis. This potentially allows us to replicate subtle prosodic clues to information structure, even if our understanding of them is incomplete.

2. BUILDING THE VOICE

2.1. Script Generation

The script for the pilot voice was designed by hand, then for the final voice, the experience gained was used to design a simple algorithm to generate the final script. The algorithm (see [12] for full details) makes use of a list of variables, such as airline type; a list of template-like sentence types with variables in the slot positions; and a method to efficiently combine the two. Each variable has a list of possible values that appear in the same phonetic and prosodic context. For this reason, some lists may have completely identical members, if the variables they are associated with provide the same information in different phonetic or prosodic environments.

The algorithm begins by rotating through all the sentence type templates. Each type is realised at least once, to ensure full coverage of sentence types. Sentence types with directly adjacent variables require every variable in the first list to be recorded next to every variable in the second. If a combination of adjacent variables is only used in one sentence type, that sentence type is generated until every combination of the two variables has been realised. Similarly, if a variable is only used in a single sentence type, there must be at least enough instances of that type to have one copy of each member of the list. Any sentence types containing variables which only appear in that one type are generated until all members of the list are realised. After that, if there are any unfinished combos (adjacent variables with unrealised combinations), the sentence type with the most variables related to unfinished combos is generated. Then, if there are any unfinished lists (lists with unrealised members), the sentence type with the most variables related to unfinished lists is generated. This guarantees that the smallest number of prompt sentences is generated, as each generated sentence will involve the realisation of the greatest number of list members.

This method worked well in producing a suitable script for a large vocabulary system, with two minor problems. Firstly, no account was taken to ensure the generated sentences were semantically or pragmatically meaningful. For example, it would generate sentences like *There is a direct flight on KLM Airlines, but you'd have to connect in Paris*. Secondly, the large number of airport names meant that the script became excessively large if each airport name was recorded in each distinct context. As a result, we split the set of airports into higher and lower priority ones, and limited the contexts in which we recorded the latter ones. Our plan is to back-off to a general purpose unit selection voice to synthesise airport names that were not recorded in the all necessary contexts.

2.2. Prosodic Diversity

To an extent, natural intonation can be expected without any explicit specification when there are only a few candidates for each phone, as long stretches of speech are likely to be taken directly from the database. As vocabulary size and complexity of context increases though, the likelihood of finding appropriately intonated units without explicit marking can be expected to decrease.

In designing the script for the pilot voice, we initially focused on the point in the dialogue where the FLIGHTS system describes the best available flights, after gathering the details of the user query. When we examined the resulting set of 'descriptive' sentences, we found that while the same words often occurred in different prosodic contexts, they nearly always occurred in different phonetic contexts as well. In part this was because the script was put together manually, with only basic coverage of the range of possible descriptive sentences; the final voice, with the automatically generated script, covers a much broader range of sentence types.

Given the limited variety of descriptive sentences in the pilot script, we reasoned that even without explicit marking, existing limited domain synthesis techniques might yield natural intonation. Consequently, to better test the approach, we added sets of sentences with the same words but different intonation and information structures. These sentences are potentially relevant to a later point in the dialogue, where the system responds to the user's request for further or clarificatory information about the flights under discussion. The most appropriate intonation for these 'clarification' sentences depends on the context established by the question. The contexts and resulting intonation and information structures are described in detail in Fig. 2.

2.3. Eliciting Contextually Appropriate Intonation

A number of potential speakers were auditioned, and the one which was best able to take into account the contexts of the sentences was chosen. Ideally, to elicit appropriate intonation, we could have recorded complete dialogues, with the speaker acting in the role of travel agent. However, given the number of sentences to record, it only seemed practical to record question and answer pairs. With the pilot

Type 1

Q: Which flight leaves Edinburgh at eleven a.m.?

A: (The KLM Airlines flight)$_{Rh}$
H* L
(leaves Edinburgh at eleven a.m.)$_{Th}$
L+H* L+H* L+H* LH%

Type 2

Q: When does the KLM Airlines flight leave Edinburgh?

A: (The KLM Airlines flight leaves Edinburgh)$_{Th}$
L+H* L+H* LH%
(at eleven a.m.)$_{Rh}$
H* H* LL%

Type 3

Q: Which airport does the KLM Airlines flight leave from?

A: (The KLM Airlines flight leaves)$_{Th}$
L+H* LH%
(Edinburgh)$_{Rh}$ (at eleven a.m.)$_{Rh}$
H* L H* H* LL%

Type 4

Q: Does the KLM Airlines flight arrive in or leave from Edinburgh?

A: (The KLM Airlines flight)$_{Th}$
L+H* LH%
(leaves Edinburgh)$_{Rh}$ (at eleven a.m.)$_{Rh}$
H* L H* H* LL%

Fig. 2. Types of clarification sentences. Each sentence type is shown with its intonation and information structure, and with an example question to which it is considered an appropriate response. Subscript Th and Rh denote theme and rheme, respectively.

voice, the script included clarification sentences like those in Fig. 2 consecutively. Our impression was that this presentation made it too difficult for the speaker, who had to cope with a changing context with every sentence. Consequently, with the final voice, we recorded blocks of sentences with the same information structure, but with different airlines, airports, etc. in each response.

3. EVALUATING THE PILOT VOICE

Certain words and phrases are more likely to be affected by prosodic markings than others. While different pitch accent types often overlapped with completely different lexical and therefore phonetic contexts, different boundary tones tended to have the same preceding context and a different following context. The most important example in this voice was the words *a.m.* and *p.m.* Half of the recorded examples are marked with H* LH% and the other half are marked with H* LL%. This means that when the APML is used, if the word appears at the end of a phrase in the middle of a sentence, it will have the continuation rise associated with such a position, while if it is the last word in a sentences, it will always have the drop in pitch and lengthening associated with the end of a declarative sentence. Without explicit marking, the lower join cost resulting from using an entire time unit (e.g. *three fifteen a.m.*) could lead to some inappropriate intonations at boundaries. Similarly, words like *non-stop* and *direct* have recordings labelled H* L and H* LL% for sentence medial and sentence final versions. The flip side of this situation is that the APML voice can force more joins per sentence, with its greater restrictions on possible units. If these are very noticeable, it can result in lower synthesis quality.

3.1. Methodology

For comparison purposes, we built two voices, a text voice, and an APML voice. The APML voice was built using a version of Festvox adapted to use APML mark-up as input, whereas the text (default) voice does not use the intonation information specified by the APML. We then carried out a perceptual experiment, divided into three comparisons. The first comparison was between the two voices using descriptive sentences. This was the smallest part of the evaluation, with only 6 sentences compared. The second comparison was again between the two voices, this time using sixteen clarification sentences, with four sentences for each of the four information structures in Fig. 2. The third comparison used the same sixteen clarification sentences, in this case comparing the APML voice using input with appropriate APML tags against the same voice but with input marked up with inappropriate APML tags (from a different information structure type).

Sixteen native English speakers participated in the perceptual experiment. None had any known hearing or language deficit. They were not paid for their participation. The experiment lasted approximately fifteen minutes. The participants were informed that they would hear over their headphones and see on the computer screen sets of question/answer pairs. Each pair would be presented twice. In both presentations the recording of the question would be identical, but the answers would have the same words spo-

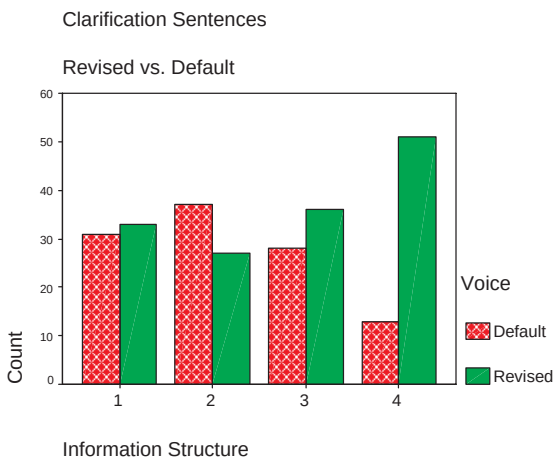

Fig. 3. Number of utterances preferred: revised (APML) vs. default (text) voice, broken down by information structure type

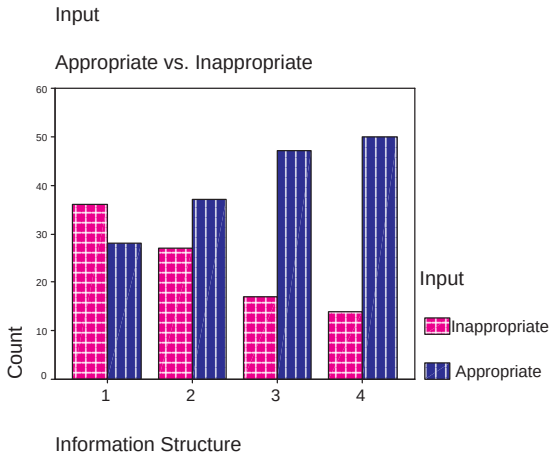

Fig. 4. Number of utterances preferred: contextually appropriate vs. inappropriate input, broken down by information structure type

ken with slightly different intonation patterns. They were instructed to listen to both version and then choose the version in which the answer sounded more appropriate given the question. Each version was presented only once.

3.2. Results

The data were analysed using three binomial tests to determine if the participants preferred one version over the other with greater frequency than would be expected by chance. If one voice was preferred in any of the experiments, individual binomial tests would be carried out for each information structure type to determine for each type whether that voice was preferred more than chance.

In the first comparison, with the descriptive sentences, the results were split evenly between the default and revised voice. This was not too surprising, given the limited variety of descriptive sentences in the pilot script.

In the second comparison (Fig. 3), the revised (APML) voice was preferred in 147 cases (57%), while the default (text) voice was preferred in 109 cases, a significant preference in favour of the APML voice ($p<.05$). When a binomial test was run for each information structure type individually, only type 4 produced results in which the revised voice was selected significantly more often than predicted by chance ($p<.001$).

In the third comparison (Fig. 4), the versions of the sentences synthesised from contextually appropriate APML files were preferred over those generated from contextually inappropriate input 162 times (63%), while the innappropriate versions were preferred 94 times, a significant preference in favour of the contextually appropriate versions ($p<.001$). When a binomial test was run for each information structure type individually, both type 3 and type 4 produced results in which the contextually appropriate versions were selected significantly more often than chance ($p<.001$).

4. DISCUSSION

The results of the clarification sentence experiments show that the addition of intonation knowledge can significantly improve the perceived appropriateness of the speech synthesis. This is encouraging, and highlights the need for further investigation as to when intonation is important and what needs to be done to get it right.

The significant effects found for only two information structure types is a useful indicator of situations in which prosodic marking has the strongest effect. This is particularly evident when looking at sentence type 4. The prosodic distinctions between a marked word in a theme and a marked word in a rheme are more subtle than the distinctions between a marked and an unmarked word. The verbs *arriving* and *leaving*, as in the sentence *The KLM flight arrives in Brussels at eleven a.m.*, are not usually marked, and therefore do not normally receive a pitch accent. But in answer to the question *Does the KLM Airlines flight arrive in or leave from Brussels?*, the marked word in the rheme (ARRIVES *in Brussels*) is the verb, leading to the strongest result for sentence type 4.

Sentence types 1 and 2 contain rheme accents at the start and end of the sentence, respectively. The subtle differences between theme and rheme accents may be overshadowed by an initial high effect or a final nuclear accent effect. Sentence type 3 if different in that the marked rheme accent is

in a place that wouldn't normally be prominent, allowing subjects to recognise it.

The lack of preference for accurate intonation could reflect a limitation in the texts used in the perception experiment. Out of context, the natural answer to the question *When does the KLM Airlines flight leave Edinburgh?* is *It leaves at eleven a.m.* or even just *At eleven a.m.*, rather than the convoluted *The KLM Airlines flight leaves Edinburgh at eleven a.m.*—the latter would only be appropriate in a context where multiple possible flights are under discussion. Had the experiment involved a lengthier dialogue to more firmly establish the context, the results might have been different.

Another possible explanation for the lack of preference for accurate intonation with sentence types 1 and 2 is that we failed to elicit sufficiently distinct intonation in the original recordings. To help assess this possibility, we carried out another, smaller perception experiment using these recordings without modification. This experiment served as a topline against which the synthesiser's performance could be compared. The topline experiment involved twelve sentences, with three sentences for each of the four information structures shown in Fig. 2. The experiment had the same structure as the comparison between the appropriate and inappropriate inputs. Four native English speakers with no known language or hearing deficits participated in the experiment.

The appropriate version was selected on 39 occasions (81% of the total), and the inappropriate version was selected on 9 occasions. A binomial test was run on the results, showing that the contextually appropriate version was preferred significantly more often ($p<.001$). Nevertheless, the fact that the inappropriate version was selected in 19% of the cases suggests that there is room for improvement in eliciting suitable intonation from the speaker.

In the topline experiment, the contextually appropriate version was chosen in a greater percentage of cases than it was in any of the tests involving the synthetic voice. The difference in performance between synthesised sentences and natural ones could result from distracting imperfections in the synthesis, or lack of continuity in the sentence level pitch contour which is not explicitly controlled for.

5. CURRENT AND FUTURE WORK

After the evaluation of the pilot voice, we built a second voice with a larger vocabulary, using the automatically generated script and a different speaker. The script was presented in dialogue form with a preceding question to elicit appropriate intonation. This voice is currently being evaluated, in conjunction with a more general unit selection voice built from the same data. Initial results suggest the method used works well, and that the intonation restriction on the unit type both improves the speed of synthesis and the naturalness of the intonation.

6. ACKNOWLEDGEMENTS

We thank Johanna Moore and Mark Steedman for their contributions towards this project. This work was supported in part by the MagiCster (IST-1999-29078), COMIC (IST-2001-32311) and FLIGHTS (EPSRC-GR/R02450/01) projects.

7. REFERENCES

[1] Alan Black and Paul Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Eurospeech '97*, 1997, vol. 2, pp. 601–604.

[2] Alan Black and Kevin Lenzo, "Limited domain synthesis," in *Proceedings of ICSLP2000*, Beijing, China, 2000.

[3] Johanna Moore, Mary Ellen Foster, Oliver Lemon, and Michael White, "Generating tailored, comparative descriptions in spoken dialogue," in *Proceedings of FLAIRS-04*, 2004.

[4] Mark Steedman, "Information structure and the syntax-phonology interface," *Linguistic Inquiry*, vol. 31, no. 4, pp. 649–689, 2000.

[5] Mark Steedman, "Information-structural semantics for English intonation," in *LSA Summer Institute Workshop on Topic and Focus, Santa Barbara July 2001*, Matt Gordon, Daniel Büring, and Chungmin Lee, Eds. Kluwer, Dordrecht, 2004, to appear.

[6] Scott Prevost and Mark Steedman, "Specifying intonation from context for speech synthesis," *Speech Communication*, vol. 15, pp. 139–153, 1994.

[7] Enric Vallduví and Maria Vilkuna, "On rheme and kontrast," in *Syntax and Semantics, Vol. 29: The Limits of Syntax*, Peter Culicover and Louise McNally, Eds., pp. 79–108. Academic Press, San Diego, CA, 1998.

[8] Michael White and Jason Baldridge, "Adapting chart realization to CCG," in *Proceedings of EWNLG9*, 2003.

[9] Michael White, "Efficient realization of coordinate structures in Combinatory Categorial Grammar," *Research on Language and Computation*, 2004, to appear.

[10] Bernadina de Carolis, Catherine Pelachaud, Isabella Poggi, and Mark Steedman, "Apml, a mark-up language for believable behavior generation," in *Life-like Characters. Tools, Affective Functions and Applications*, H. Prendinger, Ed., pp. 65–85. Springer, Berlin, 2004.

[11] L. Breiman, J. H. Friedman, J. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, London : Chapman & Hall, 1993.

[12] Rachel Elizabeth Baker, "Using unit selection to synthesise contextually appropriate intonation in limited domain synthesis," M.S. thesis, Department of Linguistics, University of Edinburgh, 2003.