Effects of Pitch Accent Type on Interpreting Information Status in Synthetic Speech

Aoju Chen and Els den Os

Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
aoju.chen@mpi.nl, els.denos@mpi.nl

10.21437/Interspeech.2005-600

Abstract

Unit selection synthesis has made it possible to produce speech with high quality. However, because it allows little control over intonation, it may produce speech with contextually inappropriate intonation. In the signalling of information status, intonation, in particular, choice of pitch accent, has been taken into account in a number of dialogue systems. Previous research shows that this can improve the perceived intonational appropriateness of synthetic speech. Using an eye-tracking paradigm, this study investigates how pitch accents H*L and L*H and deaccentuation affect the interpretation of information status in synthetic speech in English. It was found that H*L biases listeners' interpretation to new information but L*H, like deaccentuation, biases listeners' interpretation to given information. These results indicate that listeners can and do make use of intonational cues in the interpretation of information status in synthetic speech and lend strong support to the integration of intonational signalling of information status into unit selection synthesis.

1. Introduction

Unit selection [1] has made it possible to generate speech with high quality. However, it allows little control over intonation and is largely dependent on the intonation of the selected units. Consequently, it is prone to produce contextually inappropriate intonation. An important aspect of natural speech is that speakers use intonation to signal whether the current lexical entity carries new or given information. It is claimed in theories of intonational meaning [2, 3, 4, 5] that in English some pitch accents (i.e. pitch movements that take place on or start from the stressed syllables and mark the associated lexical items as prominent) convey newness whereas other pitch accents convey givenness, like deaccentuation [e.g., 6]. In an attempt to improve intonation of synthetic speech, [5] has been implemented in the open Combinatory Categorial Grammar realiser [e.g., 7] employed in a number of dialogue systems, e.g., [8, 9]. The realiser's intonational choices are implemented in the Festival synthesiser via APML [10], an XML-based markup language. In [11], listeners were presented with two versions of question-answer pairs on flight information in a synthetic voice. One version was produced using APML tags and the other was produced using no APML tags. Listeners were asked to judge for each question-answer pair in which version the answer sounded appropriate in terms of intonation. It was found that by and large answers in the APML voice were more frequently judged to be appropriate. This finding indicates that the implementation of intonational signalling of information status can significantly improve the perceived appropriateness of synthetic speech.

In this paper, we investigate whether listeners will actually make use of intonational cues in the interpretation of information status in synthetic speech, whose segmental quality may not be ideal. To this end, we examined the role of H*L, L*H (transcribed in the ToDI notation [12]), and deaccentuation in interpreting given vs. new information in a synthetic English voice.

1.1. Information structure and pitch accent type

Generally, in a conversational discourse the speaker and the listener(s) strive towards some common understanding about a particular segment of the world; the choice of pitch accent largely conveys how the speaker evaluates his contribution to the discourse. Three types of contribution have been proposed in the literature: (1) adding new information to the discourse [2, 3, 4, 5]; (2) making reference to information that is already present in the discourse [2, 3, 4, 5]; and (3) neutral, i.e. the speaker avoids to commit himself as to whether his contribution adds new information [2] or refer to given information [3].

Different theories have discussed the functions of different sets of pitch accents. Here we summarise the postulated functions of H*L and L*H, which are H* (followed by the L phrase accent) and L* (followed by the H phrase accent) respectively in [4, 5], where ToBI [13] is used to transcribe pitch contour (see [11] for a comparison between ToDI and ToBI). There is a consensus on the function of H*L but not on L*H. H*L is claimed to signal new information; L*H would seem to signal givenness according to [2, 4], newness following [5], and neither givenness nor newness according to [3].

1.2. Hypotheses

If listeners do not make use of intonational cues in synthetic speech, their interpretation of information status will not reliably reflect the functions of H*L, L*H and deaccentuation discussed above. If listeners make use of intonation cues in synthetic speech, the effects of H*L, L*H and deaccentuation can be hypothesised as follows:

Hypothesis 1: Both H*L and L*H trigger the interpretation of newness; deaccentuation triggers the interpretation of givenness;

Hypothesis 2: H*L triggers the interpretation of newness but L*H and deaccentuation trigger the interpretation of givenness;

Hypothesis 3: H*L triggers the interpretation of newness and deaccentuation givenness; but L*H is compatible with neither givenness nor newness.

2. Method

To examine these hypotheses, we combined the eyetracking technique with the action-based version of the visual world paradigm [14], following [15]. Eye fixations were monitored as subjects followed pre-recorded instructions in a synthetic voice and moved objects displayed on a computer screen by the help of a computer mouse. Each display contained four objects and four geometric shapes, as illustrated in Figure 1.

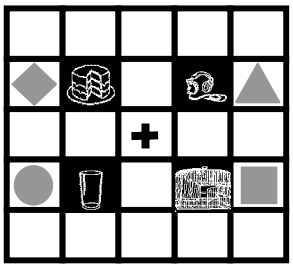

Figure 1. Example of a visual display. Geometric shapes were blue.

2.1. Experimental design

On experimental trials (vs. filler trials), two of the objects had names that shared the same stressed syllable (e.g. *candle* vs. *candy*) or the same onset-peak cluster (e.g. *cage* vs. *cake*). Each trial consisted of two consecutive instructions. The object mentioned in the second instruction was the target; its phonetically related counterpart served as the competitor. The first instruction mentioned either the target (e.g., *Put the cage below the triangle)* or the competitor (e.g., *Put the cake below the triangle*), marking the target in the second instruction either as given information or as new information (e.g. *now put the cage above the circle*). Thus, two context conditions were embedded in the first instructions: the 'given' context (where the target was mentioned) and the 'new' context (where the target was not mentioned). Because of the phonetic similarity, in the second instruction the target noun was temporarily ambiguous during the first syllable or the onset-peak cluster, and at that stage both the target and competitor noun were potential candidates for selection. The intonation of the first instruction was the same throughout the experiment; the intonation of the second instruction was varied by having the target noun said with H*L, L*H and deaccentuation, as illustrated in (1). When composing the stimuli, the target noun was also said with L*H L and H*L H to add more intonational variation to the stimuli. The effects of L*H L and H*L H were not examined in the present study because they cannot be reliably generated by our synthesiser when the sonorant material of the stressed syllable is sparse.

(1) a. First instruction: Put the { cake / cage } above the triangle;
H*L H*L H%

b. Second instruction: now put the cage below the circle.
H*L { **H*L** L% / **L*H** H% / **Deaccented** } H*L H*LH%

2.2. Predictions

Using the patterns of eye fixations to the target picture from the target word onset during the second instruction as indicators to how intonation affect the interpretation of information status, we arrived at the following predictions:

(1) Pitch accent conditions conveying newness will trigger a larger proportion of fixations to the target when it is not previously mentioned than when it is previously mentioned;
(2) Pitch accent conditions conveying givenness will trigger a larger proportion of fixations to the target when it is previously mentioned than when it is not previously mentioned;

2.3 Materials

Twenty pairs of nouns that are phonetically similar were selected from the materials used in [15]. The words were monosyllabic in twelve pairs and disyllabic in the other eight pairs. One member of each pair was assigned the role of target, the other the role of competitor. The mean lexical frequencies of the targets and competitors were identical [15]. Each of the 20 target-competitor pairs was associated with two distractor nouns, resulting in four pictures on each display (see Figure 1). Two target-competitor pairs were assigned to each of the ten conditions (2 context × 5 pitch accent conditions). Ten lists of experimental trials were constructed by varying in which of the ten conditions every two target-competitor pairs were presented. In addition to the 20 experimental trials, 48 filler trials were included to prevent subjects from developing the expectation that pictures with phonetically similar names were likely to be the targets. The 272 (20 × 4 + 48 × 4) pictures were selected from [16] and the MPI picture database. All were black and white line drawings.

In its current state, the Festival unit selection synthesiser only allows control over intonation in restricted domains, e.g., bathroom design and flight information. The spoken instructions were thus generated with the Festival diphone synthesiser, which can implement intonation choices via APML tags independently of domains. Figure 2 shows the f_0 tracks for *now put the window below the circle* with the target word *window* said with H*L, L*H, and deaccentuation.

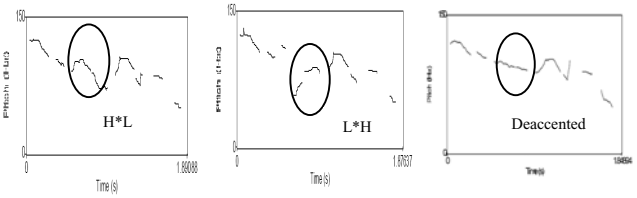

Figure 2. f_0 tracks for *now put the window below the circle* with *window* said with H*L, L*H, and deaccentuation.

2.4 Procedures

Twenty undergraduates and two postgraduates from the School of Psychology at the University of Birmingham participated in the experiment. They received either course credits or a small fee for their participation.

Subjects were tested individually. They were seated at a comfortable distance from the computer screen in a quiet room. The eyetracker was mounted and calibrated. Eye movements were monitored with a portable SR Eyelink eye-tracking system. Spoken instructions were presented to the subjects through

headphones. The structure of a trial was as follows: first, a central fixation point appeared on the screen for 500 ms. Then, a 5 × 5 grid with four pictures and four geometric shapes appeared on the screen, as the auditory presentation of an instruction was initiated. Prior to the experiment, subjects were instructed to move the object mentioned in the instruction above or below the geometric shape using the computer mouse. The positions of the pictures were randomised across four fixed positions of the grid, while the geometric shapes appeared in fixed positions on every trial. As soon as the picture was moved, the second instruction was initiated. Once the subject completed the two instructions on a trial, the next trial began. The position of the mouse cursor on the computer screen was sampled and recorded, along with the eye-movement data. A central fixation point appeared on the screen after every five trials, which allowed automatic drift correction in the calibration.

For each of the ten stimulus lists, two orders were created. One subject was randomly assigned to each order of each stimulus list. In two cases, the eyetracking data were not properly sampled due to technical problems. A second subject was then tested.

The experiment took less than 10 minutes. At the end of the experiment, subjects were asked to judge the intelligibility of the stimuli on a 7-point scale with 1 standing for hardly intelligible and 7 very intelligible.

2.5 Coding procedure

The incompletely sampled data from two subjects and data from one subject who launched few fixations before the end of the target word were excluded from coding. Data from the other 19 subjects were coded in terms of fixations. For 18 of these subjects, data from the right eye were coded; for one of these subjects, data from the left eye were coded because of calibration problems with the right eye. On each trial, the duration of a fixation was established relative to the onset of the target word in the second instruction. Graphical analysis software SUSI performed the mapping between the position of fixations, the mouse movements, and the pictures presented on each trial, and displayed them simultaneously. Each fixation was represented by a dot associated with a number, indicating the order in which the fixations occurred. The onset and duration of fixation were specified for each fixation point.

For each experimental trial, fixations were coded from the onset of the target word in the second instruction (including closure for initial voiceless consonants) to the moment when subjects clicked on the target picture with the mouse, which was taken to reflect subjects' confident identification of the target word [17]. Fixations directed to the target picture, to the competitor picture, to the distractor pictures, and to any other location on the screen were coded. Fixations falling within the cell of the grid in which a picture was presented or on the edge of that grid were coded as pertaining to that picture.

3. Results and Discussion

The coded data from 19 subjects were further analysed. The proportion of fixations to each location (i.e. target picture, competitor picture, distractor pictures, and elsewhere) was calculated in 33-ms time intervals [14] for each condition and each subject.

Figure 3 presents the proportions of fixations (averaged across subjects) to the target picture for H*L, L*H and deaccentuation in 33-ms time intervals from 0 to 1023 ms after the onset of the target word. As the minimal latency to plan and launch a saccade is about 200 ms in tasks like visual search, fixations realised in the first 200 ms of the target word are likely to be related to speech input preceding the target word. Fixations realised in the time span 200 ms after the onset of the target word are supposed to reflect input from the target word. Because the phonetically ambiguous segments of the target words were longer than 200 ms, the effects of pitch accent were expected to be strongest in the region from 200 ms to 400 ms. In the next paragraph, we take a close look at the fixation patterns in the time span 200 ms after the target word onset for each pitch accent condition.

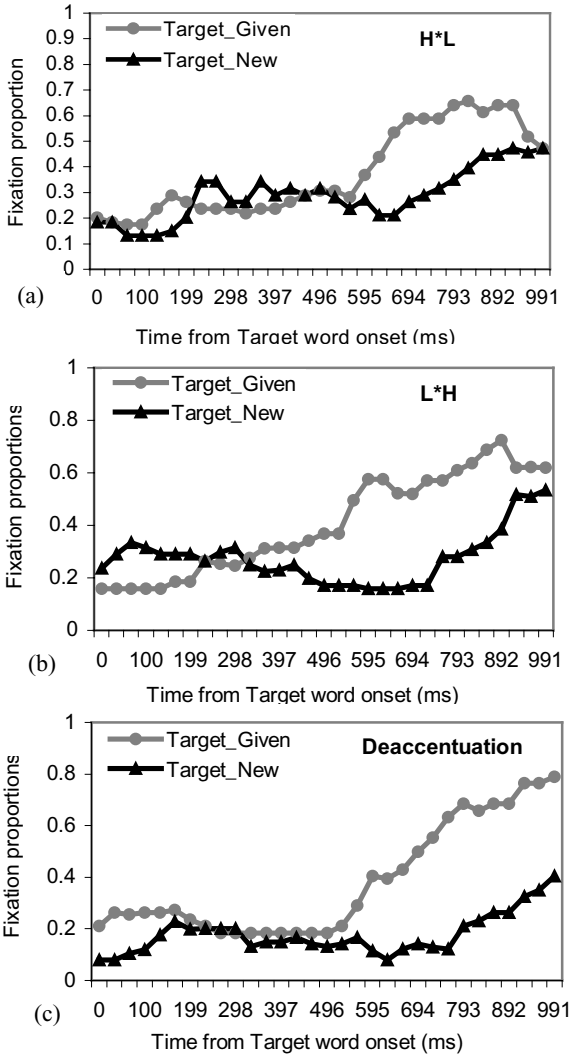

Figure 3. Fixation proportions to the target picture from the onset of the target word for (a) H*L, (b) L*H and (c) deaccentuation.

When the target word was said with H*L (Figure 3a), at about 200 ms, the proportion of fixations started to increase in the 'new-target' condition but decrease in the 'given-target' condition. Furthermore, it was relatively higher in the region 200-430 ms in the 'new-target' condition than in the 'given-target' condition. These patterns are consistent with the hypothesis that H*L conveys newness, creating a bias for the target word when it is new. When the target word was said with L*H (Figure 3b), the proportion of fixations started to

increase at about 200 ms and increased steadily till reaching 0.72 at about 890 ms in the 'given-target' condition. In contrast, in the 'new-target' condition, the proportion of fixations started to decrease at about 200 ms and continued to decrease till reaching 0.17 at about 500 ms. Importantly, these patterns are exactly what are expected on the hypothesis that L*H conveys givenness, lending support to [2, 4]. As to deaccentuation [Figure 3c], in the 'new-target' condition, a decreasing trend was present starting at about 170 ms and ending at 630 ms, whereas in the 'given-target' condition, the proportion of fixations did not change much in the region from 170-520 ms and were marginally higher than in the 'new-target' condition. This pattern is consistent with the previous finding that deaccentuation signals given information.

Interestingly, the effects of pitch accent appear to be present long after the ambiguous segment of the target word was heard. When the target word was said with L*H, the proportion of fixations to the target picture only started to increase at about 730 ms (after a steady decrease) in the 'new-target' condition while it increased steadily early on in the 'given-target' condition. Similarly, when the target word was deaccented, the proportion of fixations to the target picture started to increase much earlier (at about 500 ms) in the 'given-target' condition than in the 'new-target' condition (at about 630 ms). These observations suggest a delayed shift of attention from non-target pictures to the target picture as a result of the mismatch between the pitch accent condition and the information status of the target word. Note that there is not such an effect as regards H*L.

To evaluate the fixation patterns statistically, we conducted an ANOVA with three factors: region (0-200ms, 200-800ms), Pitch Accent Condition (H*L, L*H, Deaccented), and Information Status (given, new). There was a main effect of Region ($F_{1, 18} = 6.423$, $p < 0.05$) and a significant interaction of Region, Pitch Accent Condition and Information Status ($F_{1, 36} = 3.084$, $p < 0.05$). It is thus concluded that H*L signals new information but L*H and deaccentuation signal given information, as predicted in Hypothesis 2.

4. Conclusion

Clearly, listeners make use of intonational cues, i.e. type of pitch accent as well as deaccentuation, in the interpretation of information status in synthetic speech, in spite that the segmental quality of the synthetic speech may not be ideal (the mean intelligibility score of the stimuli is 5.8 out of 7). This lends strong support to the integration of intonational signalling of information status into unit selection synthesis.

5. Acknowledgements

This research is supported by the COMIC project (IST-2001-32311). We thank Dephine Dahan, Jan Peter de Ruiter, Barbara Schmiedtová, and Michael White for helpful comments on the experimental design, Rob Clark and Stefan Rossignol for their help with preparing the stimuli, Herbert Baumann, John Nagengast, Keren Shatzman, Johan Weustink, and Anna Zumach for their help with setting up the experiment, Antje Meyer for making it possible to conduct the experiment at the University of Birmingham, Linda Mortensen for her support during the testing and useful comments on an earlier version of the text, and Yang Luo for his help with automating the data processing procedure.

6. References

[1] Black, A. and Taylor, P., "Automatically clustering similar units for unit selection in speech synthesis", Eurospeech' 97., 2: 601-604, 1997.

[2] Brazil, D., "*Discourse intonation 1*", English Language Research, Birmingham University, 1975.

[3] Gussenhoven, C., "*On the grammar and semantics of sentence accents*", Foris, Dordrecht, 1984.

[4] Pierrehumbert, J. B. and Hirschberg, J., "The meaning of intonational contours in the interpretation of discourse", In R. R. Cohen, J. Morgen, and M. E. Pollack (eds.), *Intentions in Communication*., 271-311, MIT press, Massachusetts, 1990.

[5] Steedman, M., "Information structure and the syntax phonology interface", *Linguistic Inquiry.*, 31(4): 649–689, 2000.

[6] Cutler, A., Dahan, D., and Doneselaar, M., "Prosody in the comprehension of spoken language: a literature review", *Language and Speech*, 40: 141-201.

[7] White, M. and Baldridge. J., "Adapting chart realization to CCG", *Proceedings of EWNLG-03*, 2003.

[8] Moore, J., Foster, M. E., Lemon, O., and White, M, "Generating tailored, comparative descriptions in spoken dialogue", *Proceedings of FLAIRS-04*, 2004.

[9] Den Os, E. and Boves, L., "Towards ambient intelligence: Multimodal computers that understand our intentions", *Proceedings of eChallenges e-2003*.

[10] De Carolis, B., Pelachaud, C., Poggi, I, and Steedman, M., "Apml, a mark-up language for believable behavior generation", in H. Prendinger (ed.), *Life-like Characters. Tools, Affective Functions and Applications.*, 65–85. Springer, Berlin, 2004.

[11] Baker, R., Clark, R., and White, M., "Synthesising contextually appropriate intonation in limited domains", *Proc. of the 5th ISCA Speech Synthesis Workshop.*

[12] Gussenhoven, C., "Transcription of Dutch Intonation", In Sun-Ah Jun (ed.), *Prosodic typology and transcription: A unified approach*., 118-145, Oxford University Press, Oxford, 2004.

[13] Beckman, M.E. and Ayers, G.M. (1994), Guidelines for ToBI transcription (version 2.0, February 1994).

[14] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. and Sedivy, J. E., "Integration of visual and linguistic information in spoken language comprehension', *Science*, 268: 1632-1634, 1995.

[15] Dahan, D., Tanenhaus, M. K., and Chambers, C. G., "Accent and reference resolution in spoken-language comprehension", *Journal of Memory and Language*, 47, 292-314, 2002.

[16] Snodgrass, J. G. and Vanderwart, M., "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity", *Journal of Experimental Psychology: Human Learning*, 6: 164-215, 1980.

[17] Salverda, A, P., Dahan, D., and McQueen, J., "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension", *Cognition*, 90: 51-98, 2003.