Analysis and Modelling of Question Intonation in American English

Dmitry Sityaev, Tina Burrows, Peter Jackson, Katherine Knill

Speech Technology Group, Toshiba Research Europe Ltd.
Cambridge Research Laboratory, 1 Guildhall Street, Cambridge CB2 3NH, UK
dmitry.sityaev@crl.toshiba.co.uk

Abstract

This paper addresses the modelling in text-to-speech of the rising intonation pattern in American English which is often found in yes-no questions. A small corpus containing yes-no questions was recorded and analysed. F0 was then modelled using an automatic procedure. The paper also reports on the stability of alignment of F0 targets in rising intonation patterns.

1. Introduction

The modelling of intonation in TTS has received a considerable amount of attention in the last two decades. Various models have been developed and applied to different languages, e.g. the Fujisaki model [1], IPO model [2], etc. Data driven methods have become extremely widespread in creating statistical models for prosody, however the success of such statistical models is dependent on the availability of prosodically annotated data.

Currently, the Toshiba American English TTS system uses the closed-loop training (CLT) method described in [3], to create a codebook-based model of the F0 contour [4]. The codebook entries are trained from a corpus and applied over the domain of a prosodic word. This method is currently used for synthesising intonational patterns for declarative sentences. To model question intonation, hand-written linguistic rules are applied to the utterance-final word to achieve an interrogative F0 pattern.

The modelling of questions in TTS can also be done using a corpus-based approach. A corpus-based method for modelling F0 in questions is more robust and efficient than a rule-based method. There is no need to spend time analysing data and deriving a set of hand-crafted rules – a process which is time-consuming and may miss important generalisations. The learning of F0 patterns is done automatically, and precise points of F0 alignment are extracted directly from real speech. However, there are certain limitations here. Clark mentions that for general F0 modelling, "suitable corpora usually have a skewed distribution of pitch events: H* and L-L% are particularly frequent; L*+H and H-H% are particularly infrequent" [5]. Such a skewed distribution presents a problem for modelling questions since the low-rising pattern (L*H-H%) was found to be the most frequently occurring tune in positive (i.e. without the negative particle *not*) yes-no questions in the CallHome corpus of American English [6].

A goal of this paper is to examine whether the corpus-based F0 model can be extended to model yes-no questions in American English, thus replacing the rule-based method for modelling F0 in questions. The alignment of L* pitch accents with the segmental material in yes-no questions is also investigated. While there has been a considerable amount of research about the alignment of peaks in H* and L+H* accents in English, very little attention has been dedicated to the analysis of the alignment of valleys associated with L* pitch accents.

The structure of the paper is as follows. Section 2 describes the corpus design and labelling procedure. Section 3 provides analysis of pitch accent distribution in yes-no questions. In Section 4, the details of the acoustic analysis of yes-no question intonation patterns are presented with a focus on the alignment of L*. Section 5 provides a description of F0 modelling from a corpus of yes-no questions, while Section 6 reports on the subjective evaluation of the trained model. Finally, Section 7 provides some conclusions and highlights directions for future research.

2. Corpus design and mark-up

A US English corpus was recorded by a professional voice talent, which was aimed at covering units of speech in different prosodic contexts. Part of the corpus was specifically designed for the study and modelling of intonation in yes-no questions in American English. 100 yes-no interrogatives were created where a pitch accent was expected (a) on the last word in the sentence, or (b) on the penultimate word in the sentence.

Phonemic labelling was performed on the data using an automatic labelling procedure. Segmental labels were manually adjusted and corrected where necessary. The data were then marked up for pitch events using the ToBI transcription [7] by trained labellers.

With respect to marking points of F0 minima, the following guidelines were used. L* pitch accent markers were normally placed on the lowest F0 point within the accented word. There were often cases where the absolute trough within the falling-rising pattern was difficult to define: sometimes F0 would fall and stay level before rising. In such cases, the label was placed at the point where F0 started to rise.

Another aspect of the mark up of the sentences was judging whether L* accents actually were such, given the generally high pitch range in these utterances. The prevailing majority of yes-no questions were marked as L* for two reasons: (a) L* is known to be frequently used in English yes-no questions, and (b) the raising of the entire pitch range is quite common in questions, therefore the baseline pitch can be much higher than it would ordinarily be.

3. Pitch accent distribution in questions

100 yes-no interrogatives were first analysed with respect to pitch accent distribution. Pre-nuclear as well as nuclear accents were considered.

There were a total of 261 pitch accents labelled in the sub-corpus: 161 were pre-nuclear (intonation phrase non-final) pitch accents and 100 were nuclear (intonation phrase final) pitch accents. The majority of pre-nuclear pitch accents were

L+H* (44%) and H* (42%). There was also a small percentage of downstepped !H* accents (11%). The following accents were found in the remaining 3%: L*, upstepped H*, L+!H* and H+!H*.

Table 1 below shows the distribution of nuclear pitch accents as well as phrasal accents and boundary tones in the data. The vast majority of yes-no questions were realised with a low-rising pattern (L*H-H%). These findings are similar to those reported in [6]. Hedberg et al. also found low-rise to be the most frequently occurring pattern in positive yes-no questions.

tune	occurrence
L* H-H%	91%
L* L-H%	7%
H* H-H%	2%

Table 1: *Distribution of nuclear tunes in the data.*

4. Alignment of L* with segmental material

Acoustic analysis was performed on 98 yes-no questions containing the L* pitch accent, with a view to investigate: (a) L* alignment within the accented word; (b) the domain of alignment of L* accents.

In the majority (95%) of cases, the valley was found to align with the vowel within the stressed syllable of the accented word (see Fig. 1).

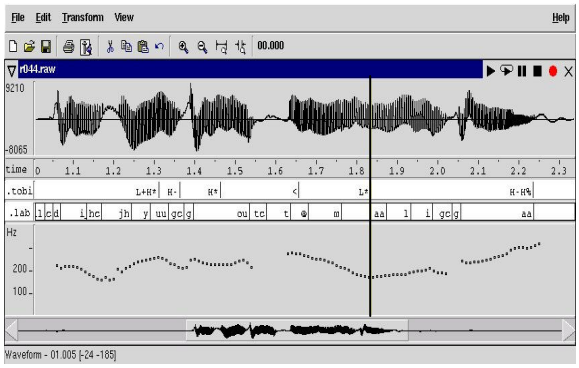

Figure 1: *An example showing L* alignment in the stressed vowel of the word "Malaga".*

In general, it was observed that sonorant onsets tended to be associated with an earlier trough alignment in the vowel. A similar finding was reported for various points of *peak* alignment in British English [8], however, van Santen and Möbius [9] argue that such an effect of onsets on peak location is "largely due to intrinsic duration differences between onsets" (sonorant vs. non-sonorant).

The domain of alignment of L* was investigated next in sentences containing one and two feet in the final accent group ("foot" is defined here in the classical Abercrombie sense [10]; "accent group" comprises a stressed syllable associated with a pitch accent plus any unaccented syllables that follow). House *et al.* [8] hypothesised that while the accent group is the phonological domain for the pitch accent association, the leftmost foot within that accent group is possibly the domain for the phonetic interpretation of that pitch accent. It was decided to test out that hypothesis.

First, the alignment of L* was investigated in accent groups containing one foot. A GLM ANOVA revealed that the number of syllables in a foot was a significant factor in determining the alignment of L* within the foot ($F(4, 42) = 12.47$, $p<0.001$): the alignment of L* into a foot was found to occur earlier in the foot as the number of syllables in the foot increased (see Fig. 2). Successive pairwise t-tests revealed that this difference was significant for all pairs of data up to 3-syllable feet. The difference in alignment was not significant between 3-syllable feet and above. For feet from 3 syllables up, the L* alignment into the foot seems to be fairly constant: around 1/5 into the foot. A similar finding was reported in [11] where the alignment of *peaks* into the foot was investigated. Klabbers and van Santen report that feet from 3 syllables up tend to have peak alignment at about 1/3 into the foot.

Similar results were found in accent groups containing two feet (the leftmost foot was the domain of pitch accent realisation). The alignment of L* into the foot was measured in the leftmost foot and compared across the feet differing in the number of syllables. A GLM ANOVA revealed that the number of syllables in the leftmost foot was also a significant factor affecting the time of L* alignment into the foot ($F(4,42) = 17.70$, $p<0.001$): as the number of syllables in the leftmost foot increased, L* would align earlier into the foot. Again, pairwise comparisons revealed similar results: the difference in alignment was found to be significant up to 3-syllable feet, and insignificant for feet of 3, 4 and 5 syllables.

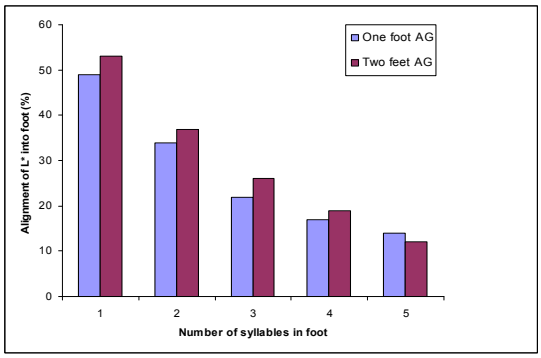

Figure 2: *L* alignment into foot expressed as percentage of total foot duration.*

Such a similarity in L* alignment is so far consistent with the prediction made in [8]. A series of pairwise t-tests were run to establish whether there is a difference between the L* alignment in accent groups made up of one foot and accent groups made up of two feet. After applying the Bonferroni adjustment, there was no significant difference found for any feet grouped by a number of syllables. This suggests that the leftmost foot in the accent group is indeed the domain of the pitch accent interpretation.

5. F0 modelling in Toshiba TTS

5.1. F0 codebook training

One of the goals of this paper was to achieve an improvement in modelling questions in the Toshiba TTS. Currently, yes-no questions are modelled by rule in the Toshiba TTS system. It was decided to adopt a corpus-based approach towards F0

modelling of interrogatives, similar to modelling F0 in declaratives.

The same 100 interrogative sentences described in Section 3 were used for training. A small subset of 9 sentences ending with a L*H-H% pattern was used for evaluation. An F0 codebook was trained on words associated with L* pitch accents only. The training method for obtaining representative contours (codebook) is described in detail in [12].

In total, there were 15 contours trained, covering different prosodic patterns. Visual inspection of the codebook revealed that the L point was always found within the stressed syllable for each prosodic pattern. This supports the findings of the acoustic analysis presented above in Section 4.

5.2. Modelling of L* pitch accent utterance-finally

The trained codebook entries for the L*H-H% pattern were incorporated into the TTS system and 5 sentences from the testing set were generated using the corpus-based version of the system (where rising intonation patterns were learned from the corpus) and the baseline version of the system (where rising intonation patterns are created by rule – modifying the last few frames of the sentence).

Perceptually, the two versions differed. The corpus-based version seemed to produce more natural utterances in terms of assigning a more appropriate stress pattern to the last pitch accented word. This is not surprising since the L target of the pattern (the point of F0 minimum) in the corpus-based version would be correctly predicted on the vowel of the stressed syllable.

To quantify the results of the two versions of the system and compare them to the original speech, the same 5 sentences were synthesised using original values obtained from speech for other aspects (grammatical attributes, accent information, duration). As can be seen in Table 2, the prediction by the corpus-based version produces lower RMS error compared to the baseline system (rule-based) on the test data.

sentence ID	rule	corpus
q015	1071	863
q017	1065	840
q054	1053	1016
q058	915	744
q098	723	634

Table 2: *RMS error (Hz) between predicted and actual F0 values for rule-based and corpus-based systems.*

5.3. Modelling of rising intonation on deaccented words following the last pitch-accented word

As a next step, it was decided to extend the model to cover sentences where the word with L* pitch accent was followed by one or more deaccented words. The current implementation (baseline) by default applies a rising intonation pattern to the last word, irrespective of whether it is accented or not. A more common pattern for yes-no questions in US English is a rising pattern starting from the last pitch accented word to the end of the utterance.

As an approximation, the (continuing) rising pattern for de-accented words was modelled as a linear interpolation between two points: F0 start and F0 end (see Fig. 3 below). The value of the F0 start point was copied from the last frame in the accented word. The value of the F0 end point was calculated by assuming a minimum rise range which was set to 0.8 octaves – a value which is currently used in the Toshiba TTS system.

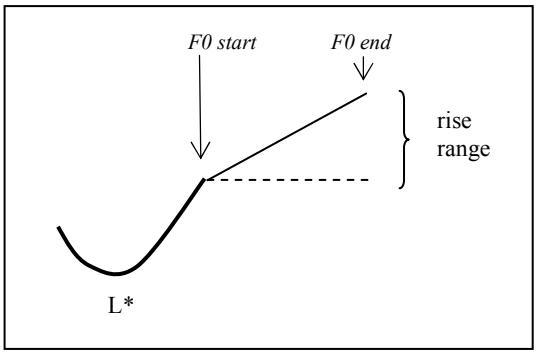

Figure 3: Schematic representation of rise interpolation.

To assess how accurate the F0 prediction is for the extended model, 4 test sentences were selected from the corpus. These sentences contained some unaccented words following the word associated with L* pitch accent and ended with H-H%. F0 values were generated using the original values obtained from speech for other aspects. It can be seen from Table 3 that the corpus-based method again produced F0 patterns with lower RMS error than the rule-based method compared against the original F0 patterns.

sentence ID	rule	corpus
q001	1120	966
q004	1019	929
q042	753	512
q076	779	781

Table 3: *RMS error (Hz) between predicted and actual F0 values for rule-based and corpus-based systems.*

6. Subjective evaluation

A perception experiment was designed to assess the naturalness of utterances synthesised with a question intonation learned from real speech.

40 yes-no questions and 40 sentences with declarative syntax intended as questions (*e.g. You want some ice in it?*) were constructed and synthesised with the question intonation using a) rule-based approach and b) corpus-based approach. The position of the nucleus was varied in the stimuli between utterance final and non-final. In addition, 20 declarative sentences were synthesised with a declarative intonation and were used as foil stimuli; they were not used in scoring. In total, there were 180 stimuli produced.

There were 3 male and 9 female native speakers of American English. Most subjects were students recruited from the University of Cambridge.

Each subject had to listen to a stimuli and respond whether the sentence sounded like a question to them by choosing the “yes” or “no” button. Reaction times were also measured from the point of the end of stimuli playback. 100 sentences were presented to each subject (40 sentences synthesised with the rule-based algorithm, 40 sentences

synthesised by the corpus-based method and 20 foil stimuli). The order of stimuli presentation was randomised each time. No single sentence was presented more than once to each subject.

Fig. 4 below presents the percentage number of responses across all subjects where subjects thought stimuli sounded like a question by sentence type. As can be seen, the corpus-based method for synthesising questions led to a higher number of correct recognitions of questions as "questions" over the rule-based method (this was significant in both cases at $p<0.001$, chi-square test). This is especially noticeable in sentences with declarative syntax intended as questions: 80% of such sentences synthesised by a corpus-driven method were recognised as questions compared to only 40% in the case of a rule-based method.

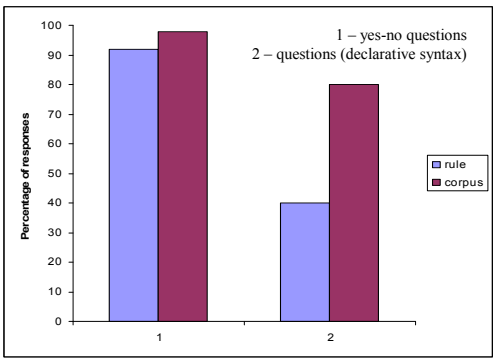

Figure 4: Percentage of sentences perceived as questions.

The subjects were also found to respond more quickly to sentences synthesised by the corpus-based system.

7. Conclusion

This paper demonstrated that a corpus-based method of learning F0 patterns can be successfully extended to model question intonation. More accurate F0 prediction was achieved using the corpus-based method compared to the original rule-based method. The synthesised speech produced also sounded more natural. This was largely due to the fact that stress patterns are captured directly from the data and points of F0 alignment are not distorted when modelling F0 question patterns from the corpus. In addition, the corpus-based method is automated and is aimed at bringing more generalisation about the way F0 is modelled in the system.

Acoustic analysis revealed that the majority of yes-no questions were characterised by L*H-H% pattern. This confirms the findings reported in [6]. L* pitch accent was found to align within the vowel of the accented syllable in the majority of cases. Moreover, it was found that the leftmost foot in the accent group was a reliable domain of pitch accent quantification.

Although there is no definite requirement on the amount of data for training a particular codebook entry, the general guideline is the more training data, the better. The distribution of ToBI events in declarative sentences is being studied with this respect. The analysis may allow more rising patterns (L*H-H% etc) to be isolated in which case it would be interesting to see whether their addition will lead to an improvement in F0 prediction. The modelling of other pitch accents is also currently under way.

8. References

[1] Fujisaki, H., 1983. "Dynamic characteristic of voice fundamental frequency in speech and singing", in P. MacNeilage (ed.), *The Production of Speech*, Springer, New York.

[2] t'Hart, J., Collier, R., and Cohen, A., 1990. *A Perceptual Study of Intonation – An Experimental-Phonetic Approach to Speech Melody,* Cambridge University Press, Cambridge.

[3] Akamine, M. and Kagoshima, T., 1998. "Analytic generation of synthesis units by closed loop training for Totally Speaker Driven Text to Speech System", in *Proceedings of ICSLP,* Sydney, Australia.

[4] Suh, C., Kagoshima, T., Morita, M., Seto, S., and Akamine, M., 1999. "Toshiba English text-to-speech synthesizer (TESS)", in *Proceedings of Eurospeech,* Budapest, Hungary.

[5] Clark, R., 2003. "Modelling pitch accents for concept-to-speech synthesis", in *Proceedings of ICPhS,* Barcelona, Spain.

[6] Hedberg, N., Sosa, J., and Fadden., L., 2004. "Meaning and configurations of questions in English", in *Proceedings of Speech Prosody,* Nara, Japan.

[7] Beckman, M. and Ayers, G., 1994. *Guidelines for ToBI Labelling,* Online MS and accompanying files. Available at "http://www.ling.ohio-state.edu/phonetics/E_ToBI".

[8] House, J., Dankovičová, J., and Huckvale, M., 1999. "Intonation modelling in Prosynth: An integrated prosodic approach to speech synthesis", in *Proceedings of ICPhS,* San Francisco, USA.

[9] van Santen, J. and Möbius, B., 1997. "Modelling pitch accent curves", in *Proceedings of ESCA Workshop on Intonation: Theory, Models and Application,* Athens, Greece.

[10] Abercrombie, D., 1967. *Elements of General Phonetics*, Edinburgh University Press, Edinburgh.

[11] Klabbers, E. and van Santen, J., 2004. "Clustering of foot-based pitch contours in expressive speech", in *Proceedings of the Fifth ISCA TTS Workshop,* Pittsburg, USA.

[12] Kagoshima, T., Morita, M., Seto, S., and Akamine, M., 1998. "An F0 control for totally speaker driven text-to-speech system", in *Proceedings of ICSLP,* Sydney, Australia.