F0 Characteristics of Yes-No Question Intonation in Arabic and English: Disambiguation Techniques for Use in ASR

Leslie Barrett
EDGAR Online, Inc., New York
lbarrett@edgar-online.com
&
Kazue Hata
Santa Barbara, California
kazue_hata@hotmail.com

Abstract

This paper presents preliminary research into the possibility of using F0 (fundamental frequency) information to enhance the performance of speech-to-speech translation engines and speech recognition software for Arabic and English. Specifically, we aim to find factors that differentiate yes-no question in both languages from other sentential types. Although previous research using cross-linguistic question data has shown F0 rise to be the main indicator of yes-no questions, the particular F0 characteristics used by listeners as perceptual cues varied. Using comparative language data, the aim of this study was to find reliable question indicators that could be detected by automated means. In an experiment with short sentences read by a native speaker of each language, we examined aspects of F0 contours in the two languages to find reliable recognition thresholds. Results indicate that reliable indicators of yes-no questions do exist for both languages and occur within the sentence-final 50 centiseconds.

1. Introduction

Prosodic cues such as F0 contours have a history of use as disambiguation tools in speech recognition. These cues have been found to disambiguate dialogue-acts [10] and cue phrases in discourse [7]. They have been shown to disambiguate even very subtle sentential features such as the scope of negation [8]. Furthermore, it is well known that prosodic values alone can be used to differentiate yes-no questions and statements in English [17,11] and other languages. This yes-no question type is characterized not only by a sentence-final F0 rise, but also an overall rise or a rise somewhere in the sentence [3, 4, 9, 12, 15, 19, 20, 21].

Both in English and Arabic, like in many other languages, question intonation is indicated by sentence-final rising contour [14]. In Arabic, the question takes an initial monosyllabic interrogative marker or uses an inverted word order, whereas in English, the question takes an auxiliary verb such as a form of 'do' or 'be'. In both languages, a yes-no question can also be realized as a declarative sentence with rising intonation.

In this paper, among multiple cues in F0 contour shape to be considered for yes-no questions, we have chosen the sentence-final F0 rise rate as the focus of investigation. Our assumption is that a certain F0 rise rate in both English and Arabic is a good indicator to disambiguate yes-no questions from other sentence types. Our study is two-fold. First, we investigated F0 rise rates along with a local factor which can have an effect on the rise, noting how F0 rise rate and rise onset location vary. Second, we attempted to determine whether it would be possible to automatically find a reliable indicator of yes-no questions. The current study was preliminary research to determine the possibility of enhancing the performance of speech-to-speech translation engines, Arabic to English, and English to Arabic, as well as Arabic speech recognition software, using F0 information. In speech-to-speech translation in particular, prosodic features stand the highest chance of success, as an indicator of sentence type, for both surviving the translation process and reducing error rates. This is due to two factors. First, background noise sometimes causes the ASR component to obscure or miss initial words. In the case of yes-no questions, this can mean the auxiliary verb is dropped, causing the identity of sentence-type to be lost. Second, the translation component can misinterpret auxiliary verbs even when the ASR output is perfect, resulting in the same effect. Thus adding a redundancy component to reinforce sentential-type information is a logical approach to improving accuracy in such cases. The current findings would enable us to identify yes-no questions automatically, given a certain threshold of the rise rate, in order to improve speech recognition in both languages.

2. Method

The domain of the speech-to-speech translation system we are dealing with is constrained in three aspects: (1) a special-domain lexicon, (2) sentential type: the most common sentential types in this system are imperatives, yes-no questions and declaratives, (3) complexity: subordination is limited to one (generally relative-clause-type) embedding per sentence.

Our data contain fifty-five question sentences in Arabic uttered by a Lebanese female speaker and forty in English uttered by an American English female speaker. These include simple yes-no questions (which are composed of three to six words) with an interrogative marker, or without an interrogative marker in inverted word order (for Arabic only) or with a rising intonation for non-inverted word order. The accent falls either on a penultimate syllable or on an antepenultimate syllable in sentence-final words, since these are common accent locations in Arabic. All questions were uttered with a rising intonation without any specific instructions to do so (the questions with a listing intonation or continuation rise were excluded from the database). For F0 analysis, we visually obtained the best-fit rise in sentence-final position by identifying the beginning and the end of the F0 rise (the best-fit version). The F0 rate was computed in semitones/centisecond (ST/cs) by dividing the F0 value difference by the duration of the rise. We used semitones in order to capture auditory perception more accurately than using the linear Hz scale. For comparison, we measured the F0 falls of a small number of declarative sentences in both languages. Since the speakers' speech rates were different, we also measured the F0 rise onset location in percent by measuring where the rise onset occurs from the end of the sentence and dividing this duration by the total duration of the sentence-final word.

3. Results

3.1. F0 rise rate and rise onset location

First, the F0 rise rates of the yes-no questions were examined in both languages. As shown in Table 1, the mean F0 rise rate in Arabic was 0.27 ST/cs (SD=0.11 ST/cs), whereas the one in English was 0.42 ST/cs (SD=0.13 ST/cs). Thus, the rise for the Arabic speaker was more gradual than it was for the English speaker, possibly indicating a language-specific difference. The difference in F0 rise rates between Arabic and English was statistically significant. (p<0.001).

language	sentence type	F0 rate in ST/cs	rise location in % from the end of the sentence
Arabic	question	0.27 (0.11)	76 (2.6)
	declarative	-0.25 (0.03)	---
English	question	0.42 (0.13)	48 (1.4)
	declarative	-0.61 (0.29)	---

Table 1*: Comparison in F0 rate between yes-no questions and declaratives in the best-fit version (The value in the parentheses shows a standard deviation)*

This tendency can be readily observed in Figures 1 and 2, which illustrate typical F0 contours and waveforms for Arabic and English yes-no questions, respectively.

The locations of the F0 rise onset are also different in these two languages. Table 1 shows that on average the rise onset occurs at 76% from the end of the sentence in Arabic, and at 48% in English. Again, Figures 1 and 2 shows this tendency: in Arabic the F0 rise onset tends to coincide with the beginning of an accented syllable, /ta:/ of /kita:'bu/ ('book'), whereas in English it tends to occur in the beginning of the final, unaccented syllable of /rInjuəl/ ('renewal'). Chahal [5] reported a similar finding in Lebanese Arabic: question intonation is characterized by a low F0 to a high rising or only a normal rising on the nuclear accented syllable, whereas declarative intonation has an F0 fall from the accented syllable toward the end of the utterance. Thus, the F0 rise in Arabic starts earlier than in English, at the accent syllable, and it takes more time to accomplish the rise, yielding a milder F0 glide.

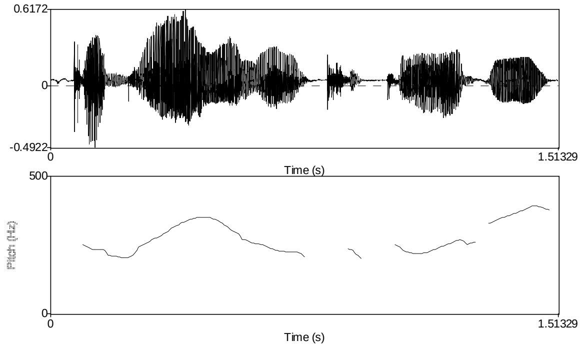

Figure 1: *F0 contour of Arabic "Is the book old*?"[qadi:'mun al kita:'bu]

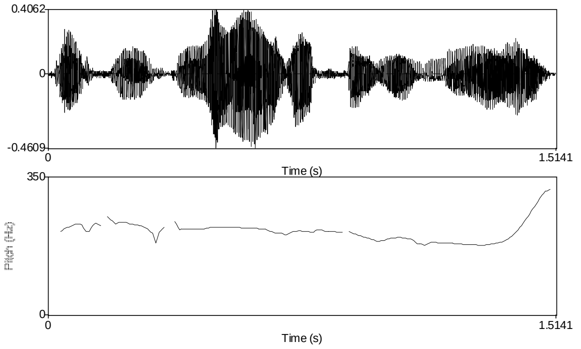

Figure 2: *F0 contour of English "Did you deal with the renewal?"*[dIdʒu dil wIδə rInjuəl]

3.2. Interaction between the rise and the accent location

The next investigation was focused on which local factor has an effect on the F0 rise rate and rise onset location. Sproat [14] reported an effect of a segmental factor on an F0 rise onset location in English monosyllabic words: in question intonation, the longer the sonorant portion, the earlier the rise onset occurs in the word. However, we suspected that F0 rise rate is chiefly influenced by accent locations in the sentence-final target word. Specifically, we believed that the determining factor for the rise rate and rise onset location is whether an accent falls on the penultimate syllable or the antepenultimate syllable.

Table 2 shows the rise rate and rise onset location for words with penultimate and antepenultimate accented syllables in both languages.

language	accent location	F0 rise rate in ST/cs	rise location in %
Arabic	penultimate	0.29 (0.11)	72 (9.7)
	antepenultimate	0.25 (0.10)	81 (11)
English	penultimate	0.46 (0.12)	46 (9.4)
	antepenultimate	0.35 (0.12)	49 (7.2)

Table 2: *The effect of accent location on F0 rise rate and rise onset location (The value in the parentheses shows a standard deviation)*

The rise onset location also shows a different trend for each of these languages. In Arabic, the onset is located 72% from the end of the sentence for the penultimate syllable but is located 81% for the antepenultimate syllable. The difference is statistically significant ($p<0.005$). In other words, when the accent falls on the antepenultimate syllable, the rise occurs closer to the beginning of the sentence-final word. In English, however, the rise onset location does not vary significantly according to accent locations ($p=0.1$), although the F0 rise starts slightly later (46% from the end of the sentence) when a penultimate syllable is accented. Thus, we could conclude that the F0 rise onset location is not particularly relevant to the location of the sentence-final accent in English but it is in Arabic. This reflects that the F0 rise coincides with the beginning of the accented syllable in Arabic, whereas the sentence-final accent is manifested as a low F0 value in our English data[1].

3.3. Towards automatic identification of the yes-no question

In this section, we demonstrate the extent to which the F0 rise rate can contribute to identifying typical yes-no questions. Can a local F0 aspect such as the sentence-final rise be used to automatically identify yes-no questions and even disambiguate them from declarative sentences? If so, does this require that a different threshold value be assigned for different languages?

To answer these questions, we computed the rate based on minimum and maximum F0 values within the F0-tracked sentence-final 50 cs (the final 50 cs version). In this method, there is no guarantee that the minimum F0 will coincide with the onset of an F0 rise and the maximum F0, with the offset of an F0 rise. Furthermore, this 50 cs might not be the final 50 cs of the sentence, since the sentence could end with voiceless segments. This method was intended to automatically extract the F0 rise rate for speech recognition without knowing the word/syllable boundary and the location of the accent.

Our aim was to examine whether this method would capture the essence of the F0 rise. Although past research [16] has shown the sentence-final stretch of about 15 cs is crucial to characterize English yes-no questions, Barrett and Hata [1] claimed that the difference between F0 rise rates obtained in the final 50 cs version (4.1 Hz/cs) and in the best-fit version (4.9 Hz/cs) was found to be below the threshold of “just noticeable difference (JND)” (i.e. JND for 4 Hz/cs is about 2.2Hz/cs); namely, that the final 50 cs would show F0 rise rates which are not perceptually different from the actual, best-fit rates (see [2, 6, 13] for the discussion of JND for F0 change).

language	sentence type	F0 rate in ST/cs	F0 rise onset location in cs
Arabic	question	0.22 (0.09)	45 (13)
	declarative	-0.27 (0.02)	---
English	question	0.34 (0.10)	29 (8.1)
	declarative	-0.26 (0.22)	---

Table 3: *Comparison in F0 rate between yes-no questions and declaratives in the final 50-cs version (The value in the parentheses shows a standard deviation)*

Table 3 shows a comparison between the F0 rise rate of yes-no questions and fall rate of declaratives in both languages based on the voiced 50cs portion of the sentence-final words, as well as the F0 rise onset location of the yes-no question. First, our results show that the F0 rise onset happens within the final 50 cs on average (45 cs for Arabic and 29 cs for English from the end of the final word). Thus, investigating English and Arabic, the sentence-final voiced 50-cs portion can be used to capture the characteristics of F0 rising. Given a certain threshold in each language, we can identify typical yes-no questions. In Arabic, this threshold value can be about 0.22 ST/cs and in English, about 0.34 ST/cs. For declarative sentences, the falling rate in both languages is about –0.26 ST/cs. This value is very different from the one in the best-fit version in English (-0.61 ST/cs). Since we took the final 50 cs, the sentence-final lowering was obscured by the maximum F0 value which occurs earlier within the 50-cs time instead of taking the exact onset of the F0 fall. The difference between questions and declaratives, however, is statistically significant ($p < 0.001$) in terms of F0 rate, both in rise and fall. Thus, taking the final 50 cs allows reasonable coverage, assuming we are dealing with a language in which the F0 rise coincides with an accented syllable. Further experiments are necessary using a new set of data for each language to support these preliminary results. We will also need to conduct tests on the threshold to determine how accurately the yes-no question can be perceptually identified.

[1] As the dataset was not controlled for sonorant portion, we cannot support either Sproat’s findings or the current accent-syllable relationship found in the Arabic data for English.

4. Conclusion

We have attempted to outline a roadmap for improving translation/speech recognition accuracy results, where sentential type is obscured, by adding redundancy to the message. This redundancy comprises the addition of prosodic features mapped on the sentence-level. Modeling sentence-type intonation is the strategy that stands the best chance of success compared with modeling the intonation of smaller phrases, given the considerations of the limited domain speech recognition.

Specifically we collected some data to find reliable question indicators that could be detected by automated means. We determined an optimal window size for gathering rise-data. Our results show that the F0 rise onset happens within the final 50 cs on average. Within that 50-centisecond window, the difference between questions and declaratives is statistically significant ($p < 0.001$) in terms of F0 rate, both in rise and fall. We noted that rise onsets for yes-no questions occur closer to the end of the sentence in English than in Arabic. However, we were not able to determine whether this crucial F0 factor used as an indicator by human listeners was related to syllable type or whether we need a single threshold of the F0 rate for both languages.

Because of the nature of the limited domain of our speech-to-speech translation system, sentence-boundary recognition is not as much of a problem as it is in other domains, where multi-sentence utterances tend to be more frequent. The question-intonation feature, because it is a sentence-level feature, could be mapped to a temporal partition of the sentence, not any syntactic or semantic constituent within it. Thus, for example, for any sentence of *n*-centiseconds intonation values can be taken at a range {*t*1..*t*n}, and that range could be calculated as a portion of the sentence *n*-cs/(*t*n-*t*1). This information could appear in XML tags (or comparable tagging convention) in the ASR output file to be ported to the translation module.

We plan to conduct further studies with more data including spontaneous speech in an effort to learn more about the crucial perceptual cues provided by F0 in English and Arabic yes-no questions.

5. References

[1] Barrett, L; Hata, K., 2002. Cues for question intonation in Arabic: Disambiguation techniques for use in automatic speech recognizer systems. *Journal of the Acoustical Society of America*, 112.5, Pt.2, 2322.

[2] Beckman, M.E., 1986. *Stress and Non-stress Accent.* Dordrecht: Foris Publications.

[3] Benkirane, T., 1998. Intonation in Western Arabic (Morocco); In Hirst and Di Cristo (eds.), *Intonation Systems: A Survey of Twenty Languages*, 345-359. Cambridge: Cambridge University Press.

[4] Bolinger, D. L., 1978. Intonation across Languages. In Greenberg, Ferguson, Moravcsik (eds.), *Universal of Human Language. Vol. 2: Phonology*, 471-524. Stanford: Stanford Univ. Press.

[5] Chahal, D., 1999. A Preliminary analysis of Lebanese Arabic intonation, *Proc. of the 1999 Conference of the Australian Linguistic Society, Australia.*

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

[7] Hirschberg, J. and Litman, D. 1993. Empirical studies on the disambiguation of cue phrases, *Computational Linguistics* 19, 501-530.

[8] Hirschberg, J. and Avesani, C. 1997. The Role of prosody in disambiguating potentially ambiguous utterances in English and Italian. In *Proceedings of ESCA Tutorial and Research Workshop on Intonation*, Athens, Greece.

[9] Hirst, D., Di Cristo, A. 1998. A Survey of intonation system. In Hirst and Di Cristo (eds.), *Intonation Systems: A Survey of Twenty Languages*, 1-45. Cambridge: Cambridge University Press.

[10] Jurafsky, D., Shriberg, E., Fox, B. and Curl, T. 1998. Lexical, prosodic and syntactic cues for dialogue acts. In *Proceedings of COLING-98 Workshop on Discourse Relations and Discourse Markers*, Montreal, Quebec, Canada.

[11] Liberman, M., and Pierrehumbert, J.B. 1984. Intonational invariance. In M. Aronoff and R.T.Oehrle (eds.), *Language Sound Structure*. Cambridge: MIT Press.

[12] Majewski, W., Blasdell, R., 1969. Influence of fundamental frequency cues on the perception of some synthetic intonation contours. *Journal of the Acoustical Society of America*, 45: 450-457.

[13] Nabelek, I., Hirsh, I.J., 1969. On the discrimination of frequency transitions. *Journal of the Acoustical Society of America* 45: 1510-1519.

[14] Nasser Eldin, S. and Rajouani, A. 1999. Analysis and synthesis of interrogative intonation in Arabic. In Proceedings of ICPhS '99, San Francisco, CA, 1509-1512.

[15] Ohala, J.J., 1983. Cross-language use of pitch: An ethological view. *Phonetica*, 40: 1-18.

[16] Pell, I., 2001. Influence of emotion and focus location on prosody in matched statements and questions. *Journal of the Acoustical Society of America*, 109:1668-1680.

[17] Pierrehumbert, J. B., 1980. The Phonology and phonetics of English intonation. MIT dissertation.

[18] Sproat, R., 1998. *Multilingual Text-to-speech Synthesis: the Bell Labs Approach.* Massachusetts: Kluwer Academic Publishers.

[19] Studdert-Kennedy, M., Hadding-Koch, K., 1973. Auditory and linguistic processes in the perception of intonation contours. *Language and Speech*, 16:293-313.

[20] Ultan, R., 1978. Some general characteristics of interrogative systems. In Greenberg, Ferguson, Moravcsik (eds.), *Universal of Human Language, Vol. 4: Syntax*, 211-248. Stanford: Stanford Univ. Press.

[21] Yuan, J., Shih, C., Kochanski, G.P., 2002. Comparison of declarative interrogative intonation in Chinese, *Speech Prosody 2002*, 711-714.