A configurational approach to the prosody of topic and focus in Egyptian Arabic. Testing the importance of accent-based and utterance-based acoustic cues

Dina El Zarka\textsuperscript{1}, Barbara Schuppler\textsuperscript{2}

\textsuperscript{1}Department of Linguistics, University of Graz, Austria
\textsuperscript{2}SPSC Laboratory, Graz University of Technology, Austria
dina.elzarka@uni-graz.at, b.schuppler@tugraz.at

Abstract
Recent research on Egyptian Arabic suggests that topic and focus are realized by different pitch events and that a focus is followed by post-focus attenuation. It is, however, not known how well these structures are discriminated and which are the relevant acoustic cues. We present results from a Random Forest based discrimination between two topic and two focus types, using 62 acoustic cues over the utterance, of which 37 are related to the target accent. All categories are discriminated very well with some confusion between the subtypes. The two most important cues are the velocity of the fall of the target accent and the scaling of the following peak, followed by the amount of fall in the accent and the intensity of the utterance-final word/syllable. Further important findings are that f0 is especially important to distinguish between target accents and to indicate register lowering, that intensity features are only related to the target word and the last word, and that duration is only relevant in the target accent. This suggests that the target accent is signaled by all acoustic cues, but primarily by a pitch fall, and post-focus attenuation by immediate register lowering and a salient intensity drop at the end of the utterance.

Index Terms: acoustic features, Random Forest, topic, focus, Egyptian Arabic

1. Introduction
One of the most investigated functions of prosody is information structure (IS), specifically focus, as the new information of an utterance \cite{2} or as indicating the presence of alternatives \cite{2}. Focus may be broad (BF) or narrow (NF). The present study is only concerned with narrow focus (NF) on a word-sized constituent. Two subtypes of NF are information focus (IF) and contrastive focus (CF) (cf. Table 1). If focus is marked prosodically, NF is assumed to be the most prominent part of the sentence \cite{3}, while given information (i.e., out-of-focus material) should be non-prominent prosodically \cite{4}. The topic of a sentence refers to what the comment is about \cite{1}. Topics may be continuous/ratified (RT), or contrastive (CT) (cf. Table 1). While the assumption that focus is most prominent is generally accepted in the literature, the claim that (contrastive) topic and focus are associated with different accent types \cite{3} is more controversial. Prosodic marking of IS has been proposed for West Germanic languages and beyond.

Prosodic studies have usually investigated only one IS category (topic or focus) at a time and the role of contrast within it (RT vs. CT, IF vs. CF) with partly controversial results. Phenetically oriented studies either investigated the topic/focus accent in isolation \cite{5, 6, 7} or studied the prominence relations between focus and background \cite{8, 9, 10}, concentrating on post-focus compression (PFC; \cite{11}) or register lowering (RL; \cite{12}). An approach that simultaneously investigates the acoustic cues in multiple words of an utterance like in \cite{13}, is only rarely applied.

Some studies found that focus strength (IF < CF) was correlated with increasing prominence, which has an impact on the choice and phonetic implementation of a pitch accent \cite{14}. \cite{15} argued that focus generally raises the register and givenness lowers it.

Earlier acoustic studies of IS in Egyptian Arabic (EA), found no prosodic reflexes of the given-new distinction and only a trend for on-focus f0 boost and PFC without any effect on intensity, duration or tonal alignment \cite{16, 17}. More recent research, however, found some effect of NF vs. BF on the scaling and alignment of turning points \cite{18}, but strong interspeaker variability. A more robust effect on scaling, alignment and the duration of the stressed syllable was found for topic vs. focus by \cite{19} who interpreted the distinction in terms of different pitch events. This study also found that contrast further raised the peak of an accent in a topic, but lowered it in focus. Another study examined the acoustic cues to prominence for 5 IS categories (RT, CT, IF, CF, BF) across the utterance and found strong effects of post-focus attenuation for all cues \cite{20}.

To summarize, these studies showed clear acoustic cues to the prosodic distinction between topic-comment and focus-background, involving both a pitch accent distinction and the absence/presence of PFC/RL. However, these studies investigated accent-based and utterance-based cues separately. They identified many highly significant cues to distinguish the categories but did not assess their relative importance. Even though f0 and intensity are mostly correlated, it could well be that one of these cues has more weight than the other in indicating prominence differences. The same is true for duration. It is also not clear whether, for instance, PFC/RL induces the alignment and scaling differences in the pitch accents. Another shortcoming is that these studies only investigated target-based cues to a potential pitch accent distinction, the sole importance of which has recently been challenged (once again) \cite{21, 22}.

In the present study, we adopt a configurational approach to prosody \cite{21, 23}, considering a wide variety of acoustic cues derived from the literature (including the mentioned studies on EA) to assess the importance and localization of these cues within the utterance. This allows us to investigate three main research questions:
1) Can IS categories and their subtypes be discriminated by a Random Forest model based on acoustic cues?
2) Which are the most important cues to IS in EA, and where in the utterance are they located?
3) Do accent-based cues play a role at all, and if so, which of these are the most important cues to IS?

2. Materials and Methods

2.1 Data and speakers

The materials consist of 6 sentences in four IS conditions (IF, CF, RT, CT). The present data set has already been used in other studies on EA [18, 19, 20]. Utterances were elicited by means of a question-answer paradigm (1) and recorded in various homes in Cairo and Alexandria by means of a head-mounted Shure microphone directly to a computer. Participants were presented with the target sentence on a computer screen and listened to a pre-recorded question stimulus. An example of a target sentence (TS) and the four stimuli are shown in (1).

(1) TS \textit{samira walaASit il-\text{an}WA:r.} \textit{Samira switched on the lights.}
IF \textit{il-\text{an}wa:r kullaha mi\text{ni}walla\text{\`a}, mi\text{n} walla\text{\`a}?} \textit{The lights are all on, who switched them on?}
CF \textit{mi\text{n} f\text{\`a} hum walla\text{\`a} il-\text{an}wa:r? sami:ra walla \text{\`a}ma:n?} \textit{Who switched on the lights, Samira or Iman?}
RT \textit{sam:ra Samalit Ɂe:\text{\`a}?} \textit{What did Samira do?}
CT \textit{Ɂa:ma:n Ɂafalit Ɂe:wi sami:ra Samalit Ɂe:\text{\`a}?} \textit{Iman closed the blinds and who did Samira do?}

Each target sentence consisted of 3 words: subject, verb and object, with the subject as target word (TW). The TW was a common female name with penultimate stress. The stressed syllable always contained the long vowel /i:/ surrounded by sonorant consonants (nasals and liquids) or /h/ to avoid f0 perturbations. The 18 speakers (11 female, 7 male) were native in EA, born and raised in Cairo or Alexandria, and all but one had university-level education and had learned English, French, or German as a second language. However, the language used in everyday life was EA for all speakers. Each target sentence was read 3 (to 4) times in 4 IS conditions by 18 speakers. The RT condition is missing for one male speaker due to recording problems. Thus, the total amount of recorded data was 1350 sentences. 161 had to be excluded for various reasons, such as disfluencies or breaks to avoid effects on duration and potentially on the behavior of tones. Our final dataset consisted of 1189 tokens for (CF = 284, IF = 317, CT = 308 and RT = 280).

2.2 Acoustic features

We used 62 acoustic features, 37 of which related to the TW accent and the other 25 were distributed over the following two words of the sentence. Each sentence was manually segmented into words (x,y,z). In addition, the stressed syllable of each word (a,b,c), the stressed vowel (v) and the final syllable (s) of the TW were segmented. F0 tracks were extracted automatically, manually corrected and smoothed using \textit{mausmooth} [24]. Targets were automatically detected in \textit{R} [25] and manually corrected: T1=f0 minimum before peak in TW (H1); L1=low elbow starting the rise to H1; H1= peak in TW; L2=low elbow after H1 ending the falling movement; T2=f0 minimum between H1 and H2; L3=low elbow starting the rise to H2; H2=second peak; H3=third peak; T3=f0 minimum after H3. f0, duration and intensity were extracted in PRAAT [26]. CPP in (s) was extracted with \textit{VoiceSauce} [27]. F0 and intensity were normalized for speaker and converted to semitones, duration was normalized for speech rate. Based on the segmentations and the annotation of target points we extracted and calculated a number of absolute and relative acoustic features. To investigate the first accent, we included static and dynamic f0 features of both the rising and the falling part of the rise-fall related to the first accent. Due to lack of space we refrain from listing and explaining all 62 features and only discuss the highly ranked features in the Results Section (3.2).

2.3 Statistical analysis: Random Forest

In order to analyze how the acoustic features contribute to the distinction of the four categories RT, CT, IF and CF, we built a Random Forest (RF) classifier. In contrast to regression-based approaches, Random Forests deal well with correlating predictors, high-order interactions between the predictors, and with a combination of a small sample size and a large number of predictors. Random Forests have already some time ago been introduced to linguistics [28] and they were recently also used to rank the importance of acoustic cues for prosodic phenomena (e.g., [29] for prosodic prominence and [30] for prosodic boundaries).

We used \textit{scikit-learn} [31] to build RF classifiers with 150 estimators (i.e., number of trees in the forest), a maximum depth of 40 (i.e., maximum number of levels in each decision tree), with a minimum samples split of 7 (i.e., mini-num number of samples placed in a node before the node is split), the square root as maximum number of features considered for splitting a node (i.e., how many features each tree is randomly assigned) and Gini impurity measure for both classification and determining feature importance. The dataset was split into 80% for training and 20% for testing, resulting in 871 training samples and 218 testing samples respectively. The RF classifier was run 30 times. We first trained a classifier with all 62 features, and then retrained it with those 20 features that ranked highest in the first run. F1-score was used to evaluate the performance of the RF.

The overall importance of each feature was estimated using the package SHAP [32]. The Shapley value is calculated by taking the average of each feature’s contribution to each coalition of features and can be used to explain the prediction of an instance. High Shapley values indicate high feature importance [33]. Importances estimated with this method take interactions and collinearities into account and thus allow for interpretations of the importance of the features relative to each other.

3. Results and Discussion

3.1 How well are the IS conditions discriminated?

Overall, we received good results for distinguishing the four categories CF, RT, CF and IF with our RF classifier. When using all 62 features, a performance of F1 = 0.61 was reached, when retraining the classifier with the 20 highest ranking features the performance reduced only marginally to F1 = 0.57. All 4 categories were approximately equally well classified. Figure 1 shows the respective confusion matrix with correct predictions in the diagonal and the incorrect predictions of the different classes. The rows of the confusion matrix repre-sent
the true labels, whereas the columns depict the model’s predictions in terms of the absolute number of tokens in the test set (n = 218).

The confusion matrix (Fig. 1) shows that CT were most frequently confused with RT and vice versa, and that IF were most frequently confused with CF and vice versa. Other confusions only occurred infrequently. Both confusion pairs are as expected, as CF/IF and CT/RT are subtypes of their respective categories focus and topic. Even though the confusion between the subtypes is substantial, the classifier correctly classified more than twice as many tokens as it confused with the other subtype. This indicates that not only the categories but also the subtypes show substantial acoustical differences.

In a prior study [19], contrast was non-significant in the whole data set, but had a significant effect on peak scaling in topics and foci separately: peaks were significantly higher in CT than in RT, but lower in CF than in IF. Significant differences among the subtypes and even larger differences between the main categories were found in another study investigating a small number of utterance-level acoustic cues to prominence [20]. In sum, the results from the confusion matrix together with the earlier findings suggest clear categorical differences between the main IS categories, but also some discriminability between the subtypes.

![Figure 1: Confusion matrix of predicted vs. true label, in absolute number of tokens in the test set (N = 218).](image)

3.2 Which are the most important cues to IS categories and how are they ranked?

3.2.1 Pitch-compression/register lowering vs. pitch boost

It has been shown in many studies on various languages that a narrow focus is followed by PFC/RL, which has also been reported for EA in earlier studies [17, 20]. This is also reflected in our feature ranking. The scaling of H2 (semitone H2) is the second most important feature (Shapely value > 0.25), less important but still among the top 20 are the f0 difference between H1 and H2 (f0diff H1-H2) and the scaling of the lowest point between them (semitone T2). Taken together these rankings indicate that PFC/RL are important cues to focus in EA. While the scaling of H2 made an equal contribution to the classification of all categories (as indicated by the portions of the bar in different colors in Fig. 2), H1 contributed more to the identification of the contrastive types CF (purple) and CT (red). However, the fact that H2 scaling was ranked higher than H1 scaling suggests that post-focus RL is a stronger cue to focus structure in EA than the f0 boost of the focus accent, which has often been regarded as an important correlate of (contrastive) focus [34].

![Figure 2: Overall importance of the 20 highest ranking acoustic features given by Shapley values estimated for the Random Forest classifier.](image)

3.2.2 Intensity and Duration

Intensity resulted to be highly relevant, as almost half of the top 20 features were related to intensity (9/20), while intensity-related features only constituted 24% of the whole feature set (15/62). The effects of focus on f0 are well documented in prosodic studies. Although it is also commonly acknowledged that duration and intensity are strong acoustic correlates of prosodic prominence [35], intensity is sometimes neglected in the investigation of prosodic focus. Even if different acoustic features are considered, prosodic studies often investigate only the focused item itself. When the post-focus domain is investigated, it is usually taken as one domain. For example, [10] found significantly lower intensity values in the post-focus domain after an initial NF in Hijazi Arabic. For EA, [20] showed that a significant intensity drop occurred in the utterance-final word/syllable after an initial NF (specifically CF) in an SVO sentence, while f0 dropped immediately after the focused word. This result already gave rise to the hypothesis that the end of the utterance makes a stronger contribution to the overall pattern in terms of intensity. In the present study, this hypothesis is corroborated. Intensity features at the end of the utterance (int_max_z, int_max_c) are ranked higher than the intensity features related to the TW (int_max_a, int_max_x, int_mean_z, int_mean_s). In contrast, the relevance of the f0 drop after the focused word (semitone H2, f0diff H1-H2) is not matched by equally highly ranked intensity values for the second word. In fact, we do not find any intensity features related to this word among the top 20 features. This indicates some independence of the usually highly correlated phonetic parameters intensity and f0 [36] with a different impact on prosodic prominence. The results of [20] also yielded a
significantly lower $f_0$ at the end of the utterance after CF, but not after IF. The present analysis, however, suggests that this effect is generally weaker than the effect of the $f_0$ drop after focus and the low intensity in phrase-final position. In fact, the final low ($\text{semitone}_T$) was not ranked among the top 20 features. Another study [37] showed that speakers apply different strategies in the prosodic marking of IS. Whereas only some speakers had very low phrase-final $f_0$ in NF compared with the other IS conditions, others realized the utterance-final syllable equally low in all conditions. These low $f_0$ levels were, however, accompanied by a strong drop in intensity only in an utterance containing NF. Our feature ranking and the findings of [20] together suggest that loudness is an important cue to IS, specifically at the end of the utterance. It would be interesting to see whether this would also hold for longer utterances or if a strong drop in intensity occurs earlier in longer utterances. Similarly, it remains to be seen whether these results can be replicated when the narrow focus is in penultimate position, or if the effect is generally much weaker in penultimate focus, as [10] showed for Hijazi Arabic. In our data, the feature ranking suggests that intensity and $f_0$ have different contributions to create prominence differences between focus and background in EA.

The only durational features among the top 20 relate to the first accent and will be discussed in the next section.

3.2.2 Target word and target accent

12 of the top 20 features are related to the TW and thus the accent associated with it. Interestingly, a dynamic pitch feature ($\text{velocity}_\text{fall}$) turned out to be the most highly ranked feature (Shapely value $> 0.30$), whereas the target-related features peak alignment ($\text{time\_align}_H1\_x\_end$) and peak scaling ($\text{semitone}_H1$) were ranked much lower (Shapely values $> 0.5\ldots 0.10$). No feature related to L2 was highly ranked. Instead, we find the dynamic velocity feature (which captures the alignment and scaling relations between H1 and L2) to be a better measure of what pitch contributes to the distinction of the categories. This difference in feature ranking is even more meaningful since both alignment features were highly significant in the accent study by [19]. While the latter study could not assess whether the earlier alignment of the targets is accompanied by a sharper falling movement, the RF analysis indicates that a sharp fall plays the most important role for the discrimination of focus, specifically CF. This is indicated by the large portion of the bar for this feature (Fig. 2, purple segment). Our results suggest that the sharp fall after the peak is what is relevant for focus marking, which is in line with claims made by [38] based on a qualitative analysis of EA and with findings by [39] for contrastive focus in English.

It should also be noted that peak alignment with respect to the end of the word was more highly ranked than other alignment measures. In addition to the more common calculation of the alignment of target points with respect to the stressed syllable and/or vowel, the word-based alignment and range features were included to test the claim by [38] that the perceptually relevant unit for tonal contours is the word as a meaningful constituent. Also, the fact that the amount of the fall within the TW ($\text{range-fall\_to\_wordend}$) was a high-ranking feature points in that direction. We may thus cautiously interpret the high importance of the features capturing slope and amount of fall between the peak and the end of the word, as well as the alignment of the peak with respect to the word boundary as an indication to the significance of the word domain for the distinction between pitch events.

Since informal observation suggested that the sharp falling movement in a focus accent often results in a final creak, we also included a measure of voice quality (CPP) in the final syllable ($i$) to test this observation. The fact that CPP is ranked among the top 20 provides some evidence for this hypothesis.

Also, the duration of the stressed syllable resulted to be important for the discrimination of focus. This result is as expected given the findings in [19]. But whereas [19] did not test syllable and vowel duration separately, the present study shows that it is predominantly the stressed vowel that has a strong effect. Finally, the duration of the entire TW was also in the top 20 list. However, it had only a low impact.

4. Conclusions

To conclude, by using a Random Forest classifier with 62 acoustic features, topic-comment and focus-background utterances could successfully be discriminated. To a lesser degree this was also true for the subtypes of topic (ratified, contrastive) and focus (informational, contrastive). The results suggest that at least the two major categories were acoustically very different in EA and that this difference was consistent across speakers. It needs to be stressed, however, that the goodness of discrimination achieved by automatic discrimination does not say much about human perception, let alone the ability to map the perceived differences to the functional categories at hand as shown by [40].

A second goal of this study was to identify the most important acoustic cues that contribute to the marking of IS categories and to assess the role played by the accent on the target word. The fact that the highest-ranked feature plus 10 other features in the top 20 were related to the first accent suggests a substantial contribution of the type of accent for the topic-focus distinction. Furthermore, we interpret the fact that only fall-related pitch features were in the top 20 as evidence for the importance of the sharp downslope to indicate focus and specifically contrastive focus, as earlier proposed by [38] for EA and [39] for English. Additionally, the maximum intensity and duration of the stressed vowel of the target word contribute to the signaling of focus, however, to a smaller degree than the $f_0$ related features.

Register lowering and post-focus attenuation in general also proved to be highly important for the distinction between all subtypes. The maximum intensity at the end of the utterance, i.e., in the final word and stressed syllable seems to be relevant for IS discrimination, specifically for the discrimination of the contrastive subtypes CF and CT.

In sum, our Random Forest based analysis supports earlier findings that topics and foci tend to be associated with different pitch events in EA. Furthermore, we were able to identify the acoustic features that jointly constitute prosodic patterns associated with the different IS categories in Egyptian Arabic, determine their distribution and assess their relative importance. Our study thereby stresses the importance of a configurational approach to prosody.

5. Acknowledgements

The work by B. Schuppler was partly funded by grant V-638-N33 from the Austrian Science Fund (FWF). We would like to thank Anneliese Kelterer and Eva Schmallegger for their support with acoustic feature extraction and scikit-learn, and Francesco Cangemi for providing an R script for automatic turning point detection.
References


