Analysis and computational modelling of Emirati Arabic intonation – A preliminary study

Abstract

This study is a preliminary investigation of intonation in Emirati Arabic (EA) (an under-researched Arabic dialect), using systematic acoustic analysis and computational modelling. First, we investigated the prosodic realisation of information focus and contrastive focus at sentence-initial, -penultimate and -final positions. The analysis of 1980 EA utterances produced by eleven EA native speakers revealed that (1) in focused words, only contrastive focus is realized with expanded excursion size, longer duration, and stronger intensity relative to their neutral focus counterparts, (2) post-focus words have a lower f_0 and weaker intensity in both contrastive focus and information focus, and (3) pre-focus words have compressed excursion size and relatively short duration. We then used computational modelling to test how much of the EA intonation could be captured by the PENTA model, with focus-defined functional categories and a number of other, putative categories. PENTAtrainer was trained on syllable-sized multi-functional targets from a subset of the production data. The model then generated f_0 contours with the learned targets and imposed them on resynthesised speech. A comparison of the model-generated f_0 contours with the natural f_0 contours showed that not only focus but also weight, stress, position of word-level stressed syllable and prosodic word are important factors determining the fine details of EA intonation. A perceptual test with native EA listeners showed that the synthetic EA f_0 contours sounded nearly as natural as the original intonation, and could convey focus nearly as accurately as natural intonation.

Keywords: Focus, PENTAtrainer, PFC, Emirati Arabic, predictive synthesis

1. Introduction

In many languages, sentence prosody encodes multiple levels of information that have effects on the surface f_0 contours, one of them being focus (Breen et al., 2010; Eady and Cooper, 1986; Rump and Collier, 1996; Xu and Xu, 2005). Previous studies have shown that focus affects not only the prosody of the word in focus (= on-focus region), but also the following words (= post-focus region), while leaving the preceding words (= pre-focus region) largely intact. By doing so, focus creates a tri-zone adjustment in the surface f_0 (Cooper et al., 1985; Eady and Cooper, 1986; Eady et al., 1986; Breen et al., 2010; Xu, 1999; Xu and Xu, 2005).

These surface f_0 variations are of interest not only to theories of intonation (Pierrehumbert, 1980; Pierrehumbert and Beckman, 1988; Nagahara, 1994; Truckenbrodt, 1995; Uechi, 1998; Grice et al., 2000; Ishihara, 2003; Sugahara, 2003; Ishihara, 2016; Kügler and Féry, 2017), but also to the computational linguistics community (Xu and Prom-on, 2014; Lee et al., 2014; Xu et al., 2015). In particular, there has been a growing interest in developing computational tools to predict fully continuous f_0 contours. This would make it possible to investigate whether the intonation of a given language can be captured successfully by a model, and could reveal aspects of intonation that cannot be easily learned from acoustic analysis alone (Xu and Prom-on, 2014; Liu et al., 2015; Xu et al., 2015). The first aim of the current study was to perform a systematic acoustic analysis of focus in EA, to find out how it shapes the surface f_0 in this dialect. Prosodic focus, however, is only one of the factors that shape surface f_0 in a language (Xu, 2005). Other factors may contribute to different aspects of intonation, but they are difficult to study using conventional experimental methods, due to the fact that they have not so far been defined as clearly as focus. Therefore, the second aim of this study was to use PENTAtrainer (Prom-on et al., 2012; Xu and Prom-on, 2014) as a modelling tool, to explore how much of the intonation of EA can be captured by specifying prosodic focus and a number of other, putative factors.[1]

1.1. Focus

1.1.1. The Notion of Focus

Focus has received many definitions in the literature, often conflicting in many respects (Rooth, 1985, 1992; Kiss, 1998; Molnár, 2002; Krifka, 2008; Zimmermann, 2008; Molnár and Winkler, 2010). However, these are predominantly theory-based, lacking solid empirical support and predictive power. Therefore, in the present paper, we adopt a working definition of focus following other empirical studies (Xu et al., 2012; Zerbian et al., 2010; Alzaidi et al., 2019). This avoids the theoretical issues with the more speculative definitions, while also reducing ambiguity in terms of the actual occurrence of focus. Specifically, we use mini-dialogues that involve WH-questions or correction-triggering statements to elicit different intonation patterns, (Eady et al., 1986; Xu, 1999; Chahal, 2001; Pell, 2001; Liu and Xu, 2005; Xu and Xu, 2005; Féry and Kügler, 2008; Wang and Xu, 2011; Alzaidi et al., 2019). Focus is therefore defined functionally, on the basis of these mini-dialogues, as illustrated in the following three examples. In (1b), *Peter* is in focus because it provides the information asked for by the WH-question (i.e. information focus). In (2b) ***Peter*** is also in focus, but the type often referred to as contrastive focus or corrective focus, because it corrects the name *John* that the speaker believes is wrong. In (3b), no word is in focus because the information asked for by the WH-question is nonspecific. Following a tradition in empirical studies (Cooper et al., 1985; Rump and Collier, 1996; Xu, 1999), we refer to this condition as neutral focus, (which is often referred to as broad focus (Ladd, 2008)). This condition serves as the baseline to which other marked intonational contours are compared. Focused words are in bold.

(1) a. Who had a book?
b. **Peter** had a book.

(2) a. Who had a book? John?
b. **Peter** had a book.

(3) a. What happened?

[1] The experimental and computational paradigms applied in the current study are not meant to capture the entirety of how focus is realised in the dialect. Nonetheless, as argued in Xu (2010), aspects of speech prosody can only be reliably examined one at a time, through strictly controlled experiments.

b. Peter had a book.

The two types of focus: information focus and contrastive focus are marked syntactically in some languages (Kiss, 1998; Vallduví and Vilkuna, 1998; Zimmermann, 2008). For example, in Modern Standard Arabic (MSA) contrastive focus is realized ex-situ (i.e., at the left periphery of the clause), while information focus is realized in-situ (Moutaouakil, 1989; Ouhalla, 1997).[2] However, other languages do not mark these two types in syntax, such as modern Arabic dialects including Hijazi (Alzaidi et al., 2019) and Emirati Arabic (as will be shown in §1.3). The in-situ syntactic position of both information focus and contrastive focus in those languages raises an interesting question of whether these two types of focus are marked differently in terms of prosodic realization.

1.1.2. Focus Prosody

Starting from the working definition of focus stated above, previous empirical studies have discovered several major prosodic characteristics of focus in various languages, which are highly consistent across speakers and repetitions (Breen et al., 2010; Chen and Gussenhoven, 2008; Eady and Cooper, 1986; Wang and Xu, 2011; Wang et al., 2018; Xu, 1999; Xu and Xu, 2005). The general findings are as follows: (i) the on-focus region is realized with expanded pitch range (often in the form of increased f_0 height), stronger intensity, and longer duration than its neutral-focus counterpart; (ii) the post-focus region is realized with compressed f_0 (hence, post-focus compression–PFC), and sometimes shorter duration and weaker intensity; and (iii) the pre-focus region is largely intact.

It has also been found that the on-focus effect is not always significant across languages and dialects that mark focus prosodically. For example, it is not observed in Hindi (Harnsberger, 1994; Patil et al., 2008), Taiyuan (Wenjun and Yuan, 2015) and Boro (Mahanta et al., 2016). In contrast, post-focus compression (henceforth PFC) is a much more reliable feature across focus-marking languages, e.g., English (Cooper et al., 1985; Eady and Cooper, 1986; Xu and Xu, 2005), Swedish (Bruce, 1982), German (Féry and Kügler, 2008), Beijing Mandarin (Xu, 1999), Korean (Lee and Xu, 2010), Japanese (Ishihara, 2002; Lee and Xu, 2012), Turkish (Ipek, 2011), Tibetan (Wang et al., 2012), Hindi (Patil et al., 2008) and Uygur (Wang et al., 2013). Furthermore, PFC was found to be highly effective in cuing focus perception (Rump and Collier, 1996; Botinis et al., 1999; Mixdorff, 2004; Liu and Xu, 2005; Ipek, 2011; Xu et al., 2012). However, there do exist many languages and dialects that do not exhibit PFC, including Wolof (Rialland and Robert, 2001), Taiwanese/Southern Min (Pan, 2008; Chen et al., 2009), Chichewa, Hausa and Northern Sotho (Zerbian et al., 2010), Taiwanese, Taiwan Mandarin Chen et al. (2009); Xu et al. (2012), Cantonese (Gu and Lee, 2007), Akan (Kügler and Genzel, 2012). The geographic distribution of the cross-linguistic patterns of PFC has been argued to have possible historical sources (Xu, 2011), but this is not a concern of the present study.

[2] MSA is a highly inflected language in which the subject is overtly nominative and the object is overtly accusative. These inflections make the OSV word order in MSA easy to process by the listener. However, the OSV word order in Arabic dialects presents a processing difficulty, because many of them have lost inflectional endings such as nominative and accusative, and hence fewer syntactic cues are available to help the listener. More information on this can be found in Shlonsky (1997), Brustad (2000), Aoun et al. (2009) and the references therein.

As for the prosodic encoding of contrastive focus, a mixed picture emerges from empirical studies, not always compatible with theoretical claims. For example, Sityaev and House (1819) find no significant differences between information focus and contrastive focus in English. This lack of the difference was also confirmed by perception experiments. Some studies did find certain small differences (Alzaidi et al., 2019; Greif, 2010; Sahkai et al., 2013). However, it has been suggested that these differences could be due to parallel encoding of emotional or attitudinal incredulity associated with contrastive focus that the studies were able to elicit (Alzaidi et al., 2019).

With regard to Arabic, there have been a number of studies on prosodic focus in some of the dialects such as Egyptian Arabic (Norlin, 1989; Rifaat, 2005; Hellmuth, 2006*b*, 2007, 2009; Chahal and Hellmuth, 2014; Cangemi et al., 2016; El Zarka et al., 2019, 2020; El Zarka and Hödl, 2021), Hijazi (Alzaidi et al., 2019; Alzaidi, 2021*a*,*b*), Lebanese (Chahal, 1999, 2003; Chahal and Hellmuth, 2014), Makkan Arabic (Alzaidi, 2022), Moroccan (Benkirane, 1998; Yeou et al., 2007; Burdin et al., 2015), Yemeni and Kuwaiti Arabic (Yeou et al., 2007). These studies differ in terms of the acoustic cues examined and the nature of the test materials used. Briefly, focus is encoded with an expanded pitch range (f_0 excursion) in the on-focus word, and also with a reduction of pitch range of the post-focus words (cf. Norlin 1989, Hellmuth 2006*b*, Hellmuth 2009, El Zarka et al. 2020 for Egyptian Arabic; Benkirane 1998 for Moroccan Arabic; Chahal 2001 for Lebanese Arabic; Alzaidi et al. 2019 for Hijazi Arabic). Besides the f_0 excursion, Chahal (2001) finds that there are other acoustic cues to focus in Lebanese Arabic, such as, the increase in the duration, intensity and F1/F2 in the on-focus region. In Hijazi Arabic, Alzaidi et al. (2019) find that focus is realized with higher f_0 and longer duration. In Egyptian Arabic, Hellmuth (2011) finds that duration, overall intensity and spectral tilt are not prosodic cues to focus. Recent studies observe variation in encoding focus across Arabic dialects. For example, focus in Makkan Arabic is realized only with an expanded excursion size, increasing the f_0 and strengthening the intensity of the on-focus region. Therefore, unlike Egyptian, Hijazi and Lebanese Arabic, Makkan Arabic is without post-focus compression (Alzaidi, 2022). This is not the only difference across Arabic dialects, but there are other differences in terms of the encoding of focus. One of them is that pre-focus region in Egyptian, Hijazi and Makkan Arabic is largely intact, whereas f_0 in the pre-focus region is compressed in some Lebanese Arabic utterances (Chahal, 2003; Chahal and Hellmuth, 2014; Alzaidi et al., 2019). These two differences are important observations, which suggest that future research needs to be more specific about different prosodic effects of focus and their possible pragmatic, semantic, and grammatical triggers.

A few computational-modelling studies have investigated small data sets from Arabic (Ibrahim et al., 2001; Hellmuth, 2018; Brown and Hellmuth, 2022). Ibrahim et al. (2001) developed a model to represent the global trend of f_0 in Egyptian Arabic. The model derived linear trendlines from a small set of affirmatives and interrogatives sentences, but did not generate all the local f_0 movements. Hellmuth (2018) used Generalised Additive Models to analyse interrogative intonation in several Arabic dialects, including Moroccan, Tunisian, Egyptian, Jordanian, Syrian, Iraqi, Kuwaiti, Buraimi, and Omani Arabic to find out the most typical contours used to encode interrogativity in these Arabic dialects. Brown and Hellmuth (2022) applied Y-ACCDIST-based method to classify various Arabic dialects, including Egyptian, Iraqi, Jordanian, Kuwaiti, Moroccan, Gulf (Buraimi, Oman), Syrian and Tunisian, based on prosodic contours.

The above brief overview of previous studies of focus across languages in general - and Arabic dialects in particular - illuminates two key cross-linguistic variables, which motivated the research questions of our production experiment in EA: (a) Does focus involve tri-zone adjustment in the dialect? and (b) Are there prosodic differences between information focus and contrastive focus? As regards computational modelling of intonation, the following subsection presents an overview of the PENTA model (Xu, 2005; Xu et al., 2022) and PENTAtrainer (Prom-on et al., 2012; Xu and Prom-on, 2014), which was applied in the current study to explore how much of the intonation of EA can be captured by specifying prosodic focus and a number of other, putative factors.

1.2. Computational Modelling

There exist many computational models of intonation, some of which have been applied to multiple languages, including Fujisaki model (Fujisaki, 1983), SFC model (Bailly and Holm, 2005), Tilt model (Taylor, 2000), and PENTA model (Xu, 2005; Xu et al., 2022). These models are associated with tools that can extract model parameters from speech data and use them to generate f_0 contours that can be imposed on resynthesised natural speech. There are also a number of computational models developed for Autosegmental-Metrical theory (AM) that aim to resynthesise English intonation (Pierrehumbert, 1981; Anderson et al., 1984; Beckman and Pierrehumbert, 1986). These early models, however, are not widely used as tools for applying or testing the theory. More recent computational tools associated with AM theory have been developed only for the purpose of analysis or annotation (Rosenberg, 2010; Hu et al., 2020), but not for generating f_0 contours that can be used in speech resynthesis. The present paper applies the PENTA model, which is capable of performing predictive synthesis of f_0 contours based on extracted model parameters (Xu and Prom-on, 2014), to EA intonation, with the aim to reveal aspects that are hard to determine from acoustic analysis alone. The following sections briefly present the core assumptions of PENTA and PENTAtrainer.

1.2.1. PENTA

The basic premise of PENTA is that speech is a system of encoding communicative meanings through an articulatory process. As outlined in Figure 1, PENTA assumes that intonational categories are primarily defined in terms of communicative functions (boxes on the far left). These functions are parallel to each other without a cross-functional hierarchy (e.g., a metrical structure that assigns both word stress and focus in the AM theory (Pierrehumbert, 1980; Ladd, 2008)), and they are associated with respective encoding schemes that specify the target approximation parameters (the second and third blocks from the left, respectively). Those parameters then control the articulation process which ultimately generates fully-detailed continuous prosody (the box on the far right). In short, speech prosody is assumed to be generated by encoding communicative functions through target approximation. In this way, a full repertoire of communicative functions are simultaneously realized in prosody, with all the details of the surface prosody still linked to their underlying sources. PENTA also assumes that how, and even whether a particular communicative function is prosodically encoded is language-specific, and that the exact details of each encoding scheme in a particular language have to be discovered through empirical investigations (Lee and Xu, 2010; Liu et al., 2013; Liu and Xu, 2005; Gu and Lee, 2007; Xu, 1999; Xu and Wang, 2009; Xu and Xu, 2005).

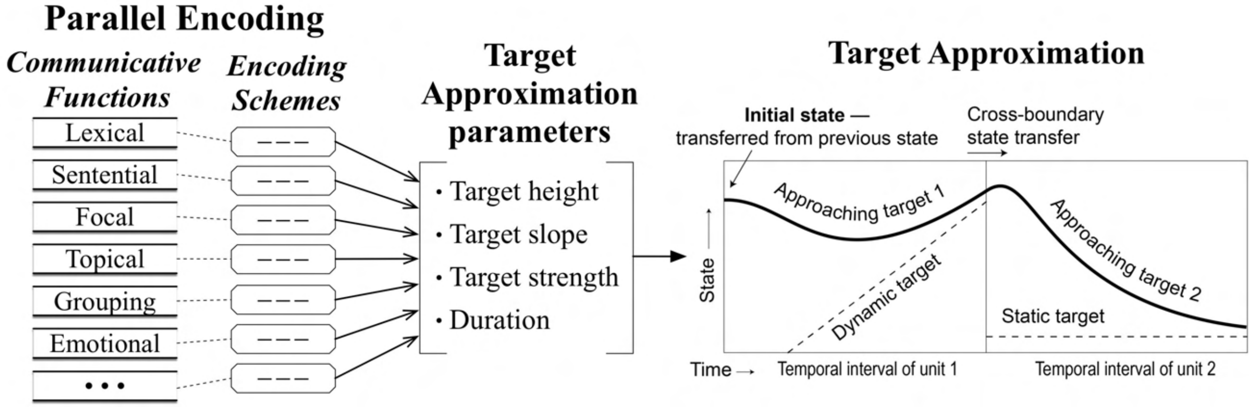


Figure 1: A schematic sketch of the PENTA model: The target approximation model of the articulation process (Xu and Liu, 2012; Xu and Wang, 2001; Xu, 2005).

Target approximation is the basic articulatory mechanism of PENTA, based on the assumption that surface f_0 contours are the results of articulatory approximation of underlying pitch target. It further assumes that pitch target approximation is synchronized with the syllable. For PENTA, therefore, establishing an encoding scheme of a functional category is to identify its contribution to syllable-sized underlying pitch targets. Following these assumptions, syllable-sized pitch targets are not exclusively associated with any single communicative function. Rather, each target is jointly determined by multiple communicative functions (Prom-on et al., 2009; Xu and Prom-on, 2014). The resulting linear sequence of multi-functional targets is what is implemented during articulation to generate the surface f_0 contour of a sentence.

1.2.2. PENTAtrainer

PENTAtrainer, written as a Praat (Boersma and Weenink, 1992) script integrated with Java programs, uses machine learning algorithms to automatically extract parameters of pitch targets from functionally-annotated speech data. The learning algorithm is simulated annealing (Kirkpatrick et al., 1983; Xu and Prom-on, 2014). As illustrated in Figure 2, there are three key elements to the application of PENTAtrainer: (a) layered functional annotation, (b) pseudo-hierarchical representation, and (c) edge-synchronization. Layered functional annotation means that each layer is annotated for a particular hypothetical function, for which the function-internal categories are putatively defined by the investigator. Note, however, that the category names only represent their identity without any phonetic specifications, as the prosodic properties reflected in the target parameters are to be learned during model training. Pseudo-hierarchical representation means that the functional layers are arbitrarily ordered. Their domains and sub-domains are combined in such a way that the layers with larger temporal intervals all project to the layer with the smallest temporal interval (syllable). In this way, no function dominates other functions. Edge-synchronization means that all the layers have fully synchronised edges with the smallest units. Therefore, the pitch target of each syllable is multi-functional, and its learned parameters carry the combined effects of all the functions present in the sentence.

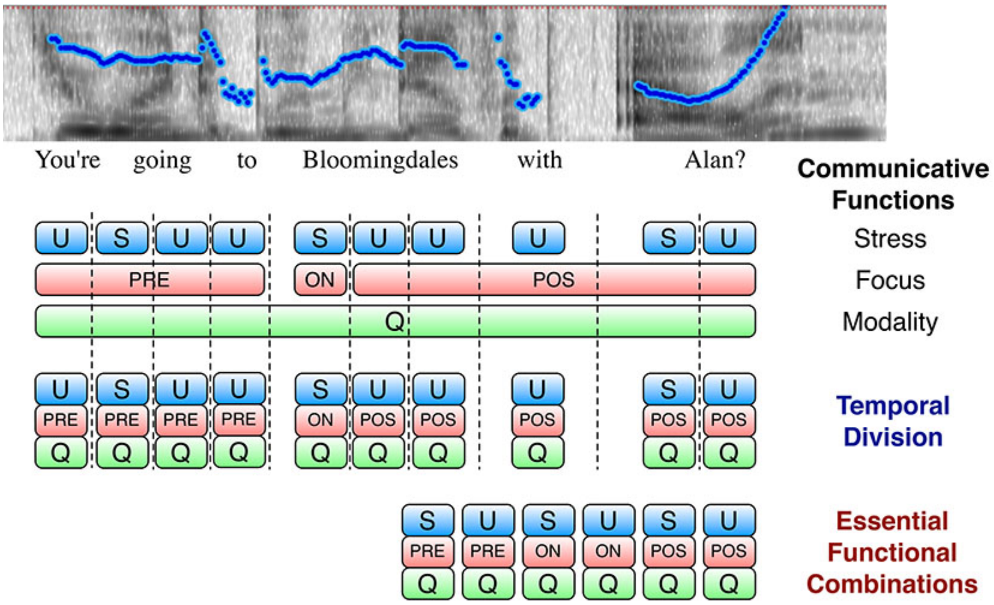


Figure 2: An illustration of how PENTAtrainer realizes functional combination through layered annotation, pseudo-hierarchical representation and edge synchronization. Here in the "Stress" layer, S denotes stressed syllables and U denotes unstressed syllables. In the "Focus" layer, PRE, ON, POS denote pre-focus, on-focus, and post-focus regions, respectively. In the "Modality" layer, Q denotes questions. The figure is adopted from Xu and Prom-on (2014), with permission.

The predicted f_0 contours are generated by the quantitative Target Approximation (qTA) model (Prom-on et al., 2009). This requires three free parameters that specify each pitch target: target slope m, target height b, and rate or strength of target approximation λ. Target slope m refers to whether the target is rising (positive) or falling (negative). Target height b indicates whether and how much the end of the target is higher (positive) or lower (negative) relative to the speaker average f_0. The rate or strength of target approximation λ indicates how rapidly a pitch target is approached. With qTA (4), given a particular pitch target, the surface f_0 contour is the result of approaching this pitch target, starting from the initial state transferred from the preceding target approximation movement, where t stands for time:

(4) $f_0(t) = (mt + b) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t}$

The transient coefficients (c_1, c_2, c_3) are calculated based on the initial f_0 dynamic state and the pitch target of current syllable. The initial f_0 dynamic state consists of initial $f_0(0)$, velocity $f'_0(0)$, and acceleration $f''_0(0)$. The dynamic state is transferred from one syllable to the next at the syllable boundary, to ensure continuity of f_0. Through the first and the second differentials, f_0 velocity and acceleration are directly carried over from the synthesized f_0 at the offset of the previous syllable. The only exception is that for the first syllable of an utterance, $f_0(0)$ is obtained directly from the original utterance, and $f'_0 0(0)$ and $f'_0(0)$ are set to 0. The three transient coefficients are composed from the following formulas (5).

(5) (a) $c_1 = f_0(0)$ - b
(b) $c_2 = f'_0(0) + c_1\lambda$ - m
(c) $c_3 = (f''_0(0) + 2c_2\lambda$ - $2c_1\lambda^2)$ / 2

Xu and Prom-on (2014) showed that it is possible to predict f_0 contours in Thai, Mandarin, and English using PENTAtrainer. Based on four functional layers (stress, focus, sentence modality, and syllable position in sentence), 78 parameters were extracted from 960 English utterances, and f_0 contours predicted by the pitch targets represented by these parameters were both perceptually and numerically close to the original intonation. Liu et al. (2015) modelled double focus intonation in American English using PENTAtrainer. Based on six functional layers (lexical stress, focus, modality, syllable position in phrase, syllable position in word, and part of speech), 146 parameters were extracted from 1960 utterances, and f_0 contours generated with these parameters were again close to those of the original sentences. In the present study, we extend the PENTAtrainer-based computational modelling to EA, with the goal of assessing the contribution of focus as well as a number of other linguistic categories often said to be present in Arabic dialects, including EA.

1.3. Emirati Arabic

Emirati Arabic (EA) is a variety of peninsular Arabic (Gulf Arabic) spoken in the United Arab Emirates, located in the East of the Arabian Gulf. It is spoken by about one million Emirati citizens. The dialect has received little attention in the research literature, although there are some recent studies on several aspects of EA, including EA verbal system (Al Kaabi and Ntelitheos, 2019), phonetic variation (Szreder and Ben-Ammar, n.d.), sluicing (Leung, 2014), subject expression and discourse embeddedness (Owens et al., 2013), acquisition (Szreder et al., 2021) as well as a comprehensive grammar (Leung et al., 2021). However, not much is known about the dialect's intonation.

There is an ongoing debate in the literature regarding whether SVO or VSO is the unmarked word order in modern Arabic dialects, including EA (Leung et al., 2021) (see also Shlonsky (1997), Mahfoudhi (2002), Alzaidi (2014), Bani Younes (2020) and the references therein). Irrespective of this debate, we take the SVO word order to provide a neutral-focus context shown in (6), to be used as the baseline to which other marked intonational contours (i.e., with focus) are compared.[3] The subscripts *IF* and *CF* stand for information focus and contrastive focus respectively.

(6) a. wiš ʔas-sālfah?
what the-story
‘What happened?’

b. Līn ḥamat Lama min Muna.
Līn protected Lama from Muna
‘Līin protected Lama from Muna.’

[3] There are cases whether VSO is not preferred, as in the EA example below, when it is uttered out of the blue.

- qabalat Lamya Hind.
  met Lamya Hind
  ‘Lamya met Hind.’ OR ’Hind qabalat Lamya.’

The VSO/VOS confusion is due to the lack of morphological inflections in Arabic dialects that encode nominative and accusative case as in Modern Standard Arabic (for more information, please see Aoun et al., 2009, and the references therein).

As in other Gulf Arabic dialects, such as Hijazi Arabic (Alzaidi et al., 2019), both information focus and contrastive focus in EA can be realised in-situ, as illustrated in (7) and (8).

(7) a. Līn ḥamat minu min Muna?
Līn protected who from Muna
'Who did Līn protect from Muna?'

b. Līn ḥamat Rīm$_{IF}$ min Muna.
Līn protected Rīm from Muna
'Līn protected Rīm from Muna.'

(8) a. Līn ḥamat minu min Muna? Nadia?
Līn protected who from Muna Nadia
'Who did Līn protect from Muna? Nadia?'

b. Līn ḥamat Rīm$_{CF}$ min Muna.
Līn protected Ri:m from Muna
'Līn protected Rīm from Muna.'

Blodgett et al.'s (2007) preliminary description of EA intonation suggests that a declarative intonation is characterized by having a pitch accent on every prosodic word (either H* or LH*), in which the high peak is associated with the stressed syllables, similar to Egyptian (Hellmuth, 2006*a*) and Hijazi Arabic (Alzaidi et al., 2019).[4] Interestingly, Blodgett et al. (2007) observed some post-focus compression, but could not determine its nature due to lack of experimental control in the study.

A relevant factor for the realisation of focus in EA is word stress, as stressed syllables have been characterised as the docking sites of pitch accents (Blodgett et al., 2007). According to Holes (1990), stress in EA is predictable, just as in other peninsular Arabic dialects. The stress assignment is determined based on (i) syllable weight, and (ii) syllable position of the word in question. In Arabic, there are three categories of syllable weight: superheavy (CVVC, CVCC), heavy (CVV, CVC), and light (CV). The stress algorithm describing the stress in peninsular Arabic is as follows:

(9) Stress is on the superheavy syllable (if any), otherwise stress is on the penultimate syllable.

This brief description of EA tells us that (i) there is no obligatory syntactic marking of information focus and contrastive focus; (ii) the primary word-stress assignment depends on

[4] In Arabic dialects including Egyptian, function words are well-established to form a prosodic unit with the following word and not with the preceding one (Al-Ani, 1992; Rifaat, 2005; Watson, 2002; Hellmuth, 2006*b*). They are not accented by default, unless they carry a focus-discourse function or inflected with the pronominalised argument of the verb. In our EA data presented in this paper, function words are all unaccented by virtue of being in neutral-focus context and also are not inflected with the pronominalised argument. Since function words are unaccented, they are assumed to procliticise to the following prosodic word rather than to the preceding one, following what has been assumed in Arabic literature (see Hellmuth, 2006*b*).

syllable weight and syllable position; and finally (iii) every prosodic word is expected to be pitch-accented (defined as local f_0 maxima). For the purpose of computational modelling, this description raises questions about how intonation can be modelled under the PENTA framework. As described in §1.2.1, PENTA assumes that the encoding schemes that shape the pitch targets are all associated with functions that are parallel to (hence independent of) each other. The predictability of stress in EA and in other Arabic dialects makes it different from lexical stress in languages like English and German, where word stress is not fully predictable. Also related to this is the technical question of whether stress, syllable weight and syllable position in a word should be all separately annotated.

1.4. Research Questions

The previous findings reviewed in §1.1.2 demonstrate that focus is prosodically encoded in some Arabic dialects, but it is not identical across dialects. More importantly, focus accounts for only part of the intonation in a language, and it is still unclear how the rest of the intonational patterns can be accounted for. As mentioned in §1.2, computational modelling can provide a means of studying aspects of intonation that are not well defined. The following are the research questions addressed in the present study:

(10) a. Does focus in Emirati Arabic involve tri–zone prosodic adjustments as found in many other languages?

b. Are there prosodic differences between information focus and contrastive focus?

c. Can Emirati Arabic focus intonation be computationally captured by an articulatory-functional model?

d. What can computational modelling tell us about Emirati Arabic intonation that we cannot easily learn from acoustic analysis alone?

2. Production Experiment

2.1. Methods

Through systematic comparisons between information focus, contrastive focus, and neutral focus, we investigated how focus and its types are prosodically encoded in EA. We elicited focus using the question-answer paradigm. We then performed detailed acoustic analyses of the elicited utterances to find the reliable acoustic cues to focus. We restricted our test materials to five-word declarative sentences spoken with information focus, contrastive focus, and neutral focus in three different sentential positions: initial, penultimate and final position. Given the shortage of space, we limited the acoustic features investigated to mean f_0, maximum f_0, minimum f_0, excursion size, duration and mean intensity.

2.1.1. Materials

The stimuli were short declarative sentences. To make extensive f_0 alignment analyses viable, we used words that have sonorant onsets and no coda consonants, wherever possible (Himmelmann and Ladd, 2008). The target sentences were in the form of /*Līn* ḥamat

Lama min *Muna*/ "Līn protected Lama from Muna", similar to the experimental paradigm used by Chahal (2001) and Alzaidi (2022), with modifications.[5] The italicized words are referred to as "key words." They varied in word length, stress pattern, syllable weight of stressed syllable, and focus, as illustrated in (11). Word length varied from monosyllabic to trisyllabic. Lexical stress varied between word-final (including monosyllabic words) and non-final. Syllable weight of stressed syllable was either light, heavy or superheavy.

Three sentence groups, illustrated in (11), were composed for examining f_0 contours at three locations in the sentence: initial, penultimate, and final. In each sentence group, the alternative words in the same location were rotated to form different sentences. Sentences in each group were produced in three focus conditions: neutral focus, information focus, and contrastive focus on the underlined word.[6] Throughout the paper, in all the EA sentences presented, the stressed syllables are in boldface, and syllable boundaries are marked with a dot.

(11) a. **Līn**/**Lī**.na/Ma.**nāl**/Ma.**lī**.kah **ḥa**.mat **La.**ma min **Mu.**na
Līn/Līna/Manāl/Malīkah protected Lama from Muna

4(words) x 3(foci) x 5(repetitions)=60

b. Līn **wa.**dat **Rīm**/**La**.ma/Na.**wāl**/May.**mū**.nah li **Mu.**na
Līn took Rīm/Lama/Nawāl/Maymūnah to Muna

4(words) x 3(foci) x 5(repetitions)=60

c. **Līn ya.**bat **La.**ma li **Mu**.na/**Dī.**ma/La.**yān**/Mu.**nī**.rah
Līn brought Lama to Muna/Dīma/Layān/Munīrah

4(words) x 3(foci) x 5(repetitions)=60

Focus is controlled by having subjects produce the target sentences as answers to prompt questions that ask about specific pieces of information available both in the target sentence and in the anecdotes that were read before. The prompt questions are shown below, together with illustration of focus locations in example target sentences.

Prompts:	*Target*:
What happened?	Lina protected Lama from Muna.
Who protected Lama from Muna?	Lina protected Lama from Muna.
Who protected Lama from Muna? Layan?	Lina protected Lama from Muna.

One anecdote at a time was shown on the computer screen for the subject to read silently. Once the subject finished reading the anecdote, they were asked to read the target sentence

[5] Chahal (2001) did not use short anecdotes preceding the question-answer pairs in her test materials. In the current study, we did use anecdotes, and followed the design of the test materials by Alzaidi et al. (2019). However, Alzaidi (2022) used short anecdotes preceding the question-answer pairs.

[6] The verbs /wadat/ and /yabat/ in the target sentence 11b and 11b, were pronounced as /wadat/ and /yabat/, respectively, by our participants who are native speakers of EA spoken in AL-Ain, and not as /waddat/ and /yaabat/ that might be possibly pronounced in connected speech in other Arabic dialects. Furthermore, the /h/ final sound in /Malīkah/ in the target sentence 11a, in /Maymūnah/ in the target sentence 11b and in /Munīrah/ in the target sentence 11c were all overtly pronounced by these speakers. The test materials, including the short anecdotes in Appendix A were all checked by a native speaker of EA, who is also a trained linguist, completing her PhD in Emirati Arabic at New York University Abu Dhabi.

as an answer to the prompt question, recorded by a female native speaker of EA. The answer and the question were seen on the screen by the subject. A sample anecdote is provided in (12) (see Appendix A for a sample of the anecdotes used).

(12) a. A sample of the type of ‘anecdotes’ in Arabic.

لين ولما ومنى أخوات. منى تبا تضرب لما ، لكن لين حمتها

b. Glossing

Līn wa Lama wa Muna ʔaḫwāt. Muna taba tiḍrib Lama lākin Līn
Līn and Lama and Muna sisters Muna wanted hit Lama but Līn
ḥamatha.
protected.her

‘Līn, Lama and Muna are sisters. Muna wanted to hit Lama, but Līn protected her.’

In order to elicit the colloquial productions and to keep the standardised register of Arabic (i.e., MSA) to a minimum following Hellmuth (2006*b*); Alzaidi et al. (2019), the EA lexical words and spelling conventions are used. Some examples of EA lexical words used that differ from MSA are shown in Table 1.

Table 1: Example of EA lexical words used in the data sets to elicit colloquial register (with their MSA equivalents).

EA	**MSA**	**Gloss**
čaif	kaif	how
kūb	ʔināʔ	cup
šu	māḏa/ma	what
taba	turīd	wanted
turūḥ	taḏhab	go
yabat	ʔaḥḏarat	brought.her
yāt	ḥaḏarat	attended.3sf[7]
ʔssalifah	ʔalqišah	story
ʔirfijāt	ʔaṣdiqā	friends
ḥālič	ḥaluk	your.situation

2.1.2. Subjects

Eleven female native speakers of Emirati Arabic participated in the production experiment.[8] They were all undergraduate students at United Arab Emirates University, born and raised in the city of Al-Ain, UAE. They had no self-reported speech and hearing disorders and their ages ranged from 18 to 23 (*mean age* = 20.8 years, *SD* = 1.7 year).

[7] 3= third, s= singular, and f= female

[8] Male subjects whom we managed to approach at the time of the experiment refused to participate due to their personal choice.

2.1.3. Recording

Recordings were conducted in the Phonetics Laboratory at the Department of Linguistics at United Arab Emirates University. Recordings were made using a Shure professional unidirectional head-worn dynamic microphone (Model WH20XLR) connected to an ASUS laptop via a USBPre 2 preamplifier (Sound Devices). The utterances were directly recorded into a computer using version 2.4.2 of Audacity(R) recording and editing software (Team, 2020) with a sampling rate of 44,100Hz and 16-bit resolution.

Materials were presented in PowerPoint, with one short anecdote per slide. After reading the projected anecdote, a question on a factual point in the anecdote with its answer were presented on another slide. Participants were asked to read the target sentence as an answer to a prompt question recorded by a female native speaker of Emirati Arabic (an undergraduate student at UAE university, born and raised in Al-Ain with no self-reported speech and hearing disorders). The participants were instructed to say the projected material in a natural way at a normal speech rate. The entire trial, including the experimenter's question, was repeated if there was any hesitation in the participant's answer. Each participant went through a number of practice trials until they were familiar with the procedure. The test materials were presented in random order, and a different order was used for each subject. Only one question-answer pair was projected at a time. The average duration of each recording session was about 35 minutes.

2.1.4. F_0 Extraction and Measurement

All the tokens (total = 1980) were first labelled manually into syllables using ProsodyPro (Xu, 2013), a script running under Praat (Boersma and Weenink, 1992). No tokens were excluded for either dysfluencies or other issues. Then, the acoustic measurements of the stressed syllables of the target words in (13) were extracted by ProsodyPro. Acoustically, we took the syllable to start from the beginning of the consonant closure (i.e., onset) and end by the end of the release of the coda consonant - or the offset of the vowel when there was no coda.

(13) a. **Max $\mathbf{F_0}$ (Hz):** highest f_0 in the stressed syllable of the key words.

b. **Min $\mathbf{F_0}$ (Hz):** lowest f_0 in the stressed syllable of the key words.

c. **Mean $\mathbf{F_0}$ (Hz):** average of all f_0 points in the stressed syllable of the key words.

d. **Excursion Size (st.):** f_0 distance in semitones between the lowest pitch and highest pitch in the stressed syllable of the key words.

e. **Intensity (dB):** mean intensity values in the stressed syllable of the key words.

f. **Duration (ms):** duration of the stressed syllable of the key words.

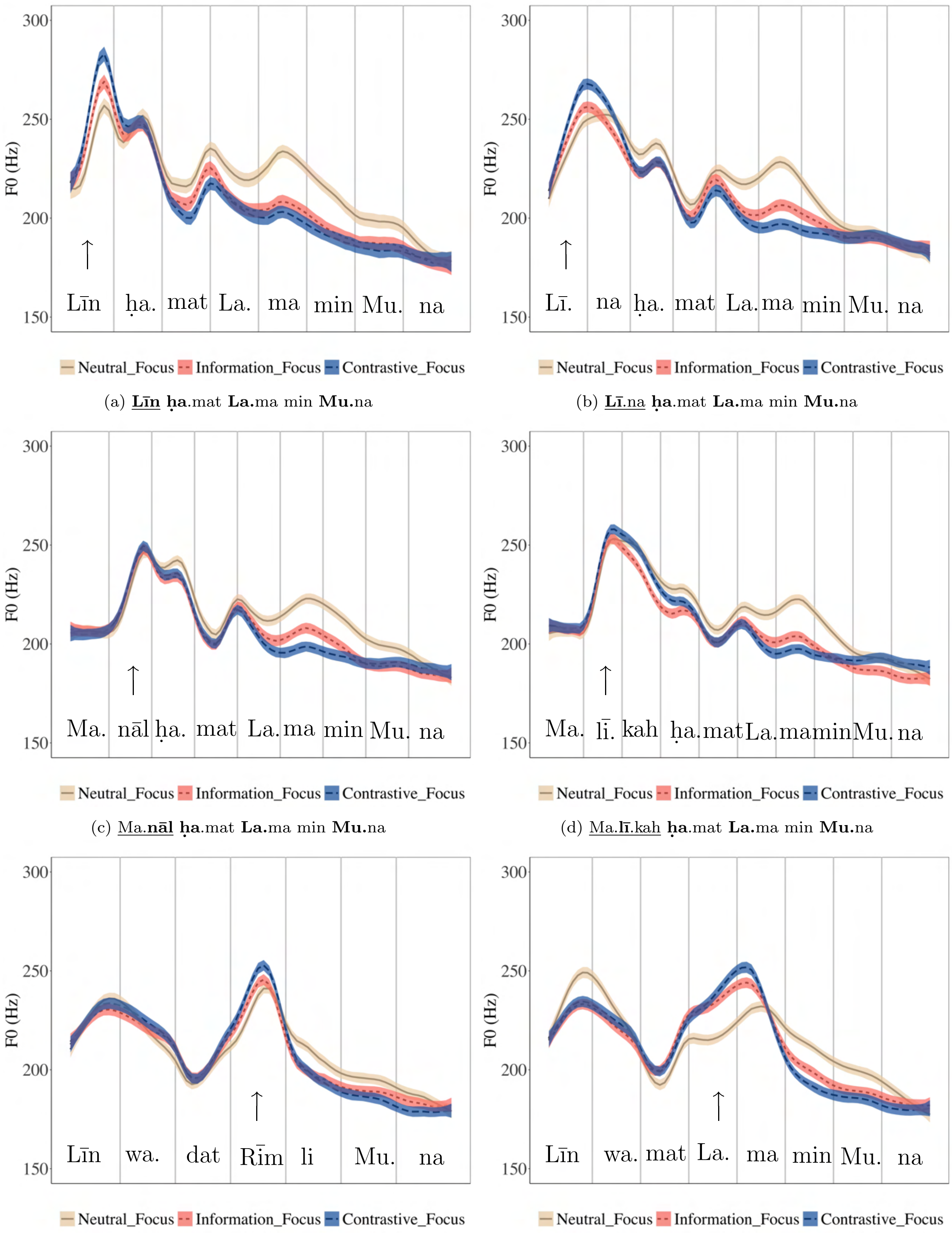
2.1.5. Effect of Focus on the Global f_0 Curve

The graphs in Figure 3 display the smoothing spline ANOVA plots of f_0 contours of the different focus conditions, each corresponding to a separate curve: neutral-focus, information focus (sentence-initial, sentence-penultimate and sentence-final), and contrastive focus (sentence-initial, sentence-penultimate and sentence-final). Using the gss package (Gu, 2014) in R (R Core Team, 2019), we applied the Smoothing Spline Analysis of Variance (SSANOVA model) to the time-normalised f_0 (10 points/syllable).[9] We included focus conditions, normalised time, and their interaction, as predictors of the dependent variable, i.e., f_0 ~ focus condition * normalised time for each sentence type. In all SSANOVA figures, f_0 means are displayed by lines and 95% confidence intervals are displayed by transparent ribbons. Where the ribbons do not overlap, the difference between their represented conditions is statistically significant.

Looking at the graphs in Figure 3, we can observe the following: First, every content word in all the target sentences with and without information focus/contrastive focus is realized with an f_0 peak near the stressed syllables of the words. This is in conformity with Blodgett et al.'s (2007) claim mentioned in §1.3, which is not yet corroborated, that every prosodic word is expected to be pitch-accented in EA intonation, where a pitch accent is defined as local f_0 maxima associated with a stressed syllable.[10] Second, the f_0 peaks fall inside the lexically stressed syllable in most of the target words. This does not differ across the focus conditions. However, the case of /lama/ in Figure 3f, 3i and 3l is different, as its highest f_0 peak is in the second syllable rather than in the first syllable which is supposed to be stressed according to the stress rule in (9). This is unlikely to be a case of peak delay due to articulatory limit (Xu, 1998), because f_0 in the first syllable drops to a local minimum before rising again to the second peak in the next syllable. There are therefore two separate f_0 movements. The reason for this apparent violation of the stress rule in (9), which has been observed before (Chahal, 2001), is not clear, and can only be studied further in future research. Third, the domains of the local f_0 maxima on the stressed syllables are very local under focus, even in the case of /lama/ discussed above. That is, f_0 starts rising from the onset of the stressed syllable until it reaches the highest point, and then starts to lower until the end of the stressed syllable, without spanning across the entire word, although in the case of /lama/, the actual peak is slightly delayed into the beginning of the next syllable. This is clearly visible in almost all the graphs in Figure 3. Fourth, f_0 peaks are higher in both information-focus and contrastive-focus words than those in the same words in the neutral-focus sentences for all sentential positions. Fifth, f_0 peaks in information-focus words and contrastive-focus words are not consistently different from each other in all sentential positions, unlike in Hijazi Arabic (Alzaidi et al., 2019). This will be further examined statistically in §2.1.6. Sixth, f_0 peaks of post-focus words are lower than those of the same words in the neutral-focus condition. Finally, f_0 peaks of post-focus words in information-focus sentences are not very different from f_0 peaks of their counterpart in contrastive-focus sentences.

[9] The time-normalised f_0 contours are used only for graphical observation and to make visual comparisons across the focus conditions. The acoustic measurements, however, were taken from the non-time-normalised f_0 tracks.

[10] Note that we are not endorsing the notion of "pitch accent" here, because the definition as given here is phonetic and descriptive rather than functional as opposed to functional concepts like tone and focus.



(a) **Līn** **ḥa**.mat **La**.ma min **Mu**.na

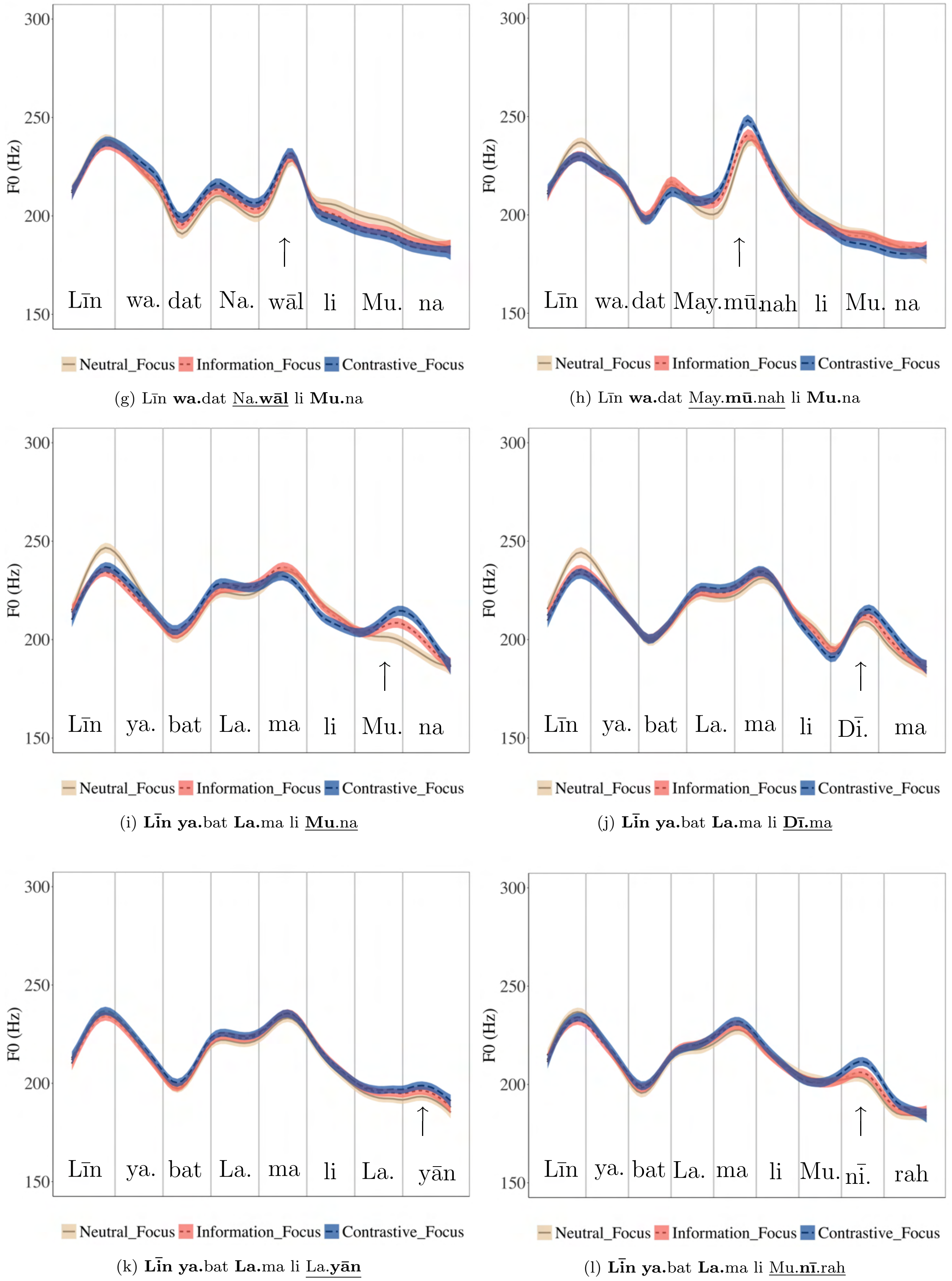
(b) **Lī**.na **ḥa**.mat **La**.ma min **Mu**.na

(c) Ma.**nāl** **ḥa**.mat **La**.ma min **Mu**.na

(d) Ma.**lī**.kah **ḥa**.mat **La**.ma min **Mu**.na

(e) Līn **wa**.dat **Rīm** li **Mu**.na

(f) Līn **wa**.dat **La**.ma li **Mu**.na



(g) Līn **wa.**dat Na.**wāl** li **Mu.**na

(h) Līn **wa.**dat May.**mū.**nah li **Mu.**na

(i) **Līn ya.**bat **La.**ma li **Mu.**na

(j) **Līn ya.**bat **La.**ma li **Dī.**ma

(k) **Līn ya.**bat **La.**ma li La.**yān**

(l) **Līn ya.**bat **La.**ma li Mu.**nī.**rah

Figure 3: SS-ANOVA plots of time-normalised f_0 contours: The lines display f_0 means and the surrounding ribbons display 95% confidence intervals. The vertical lines mark the syllable boundaries. Stressed syllables are in bold. The word in focus is underlined. The upper arrow indicates the position of the stressed syllable of on-focus words.

To verify the visual observations, a series of Linear Mixed-Effects models were performed on all the measurements listed in (13) using the lme4 package (Bates et al., 2015) in R (R Core Team, 2019). We started with a baseline model, which included by–speaker, by–sentence type[11] and speaker–by–sentence type random intercepts and slopes for focus condition (neutral-focus, information-focus, contrastive-focus) (Barr et al., 2013).[12] We then included focus condition as a potential fixed effect, if it was judged to be superior to a less fully specified model by likelihood ratio tests. The models for the three focus positions were fitted separately. *P* values were obtained by likelihood ratio tests of the models with and without the fixed effect (e.g., focus conditions). For a significant main effect, the post-hoc comparisons were conducted by the emmeans package (Searle et al., 1980) in R (R Core Team, 2019). All statistical effects will be reported at a significance level of 0.05. The R codes used are presented in Appendix B. Model prediction plots with the predicated mean and 95% confidence intervals are shown in Appendix C.

2.1.6. Quantitative Analyses

A. On-focus Region

Table 2 presents the means and standard deviations of the acoustic measurements (13) of the stressed syllable of the target word under focus. It shows that the mean values increase across the focus conditions: neutral-focus < information-focus < contrastive-focus. Across all three sentential positions, excursion size, duration and mean intensity significantly increase across all focus conditions (contrastive focus > information focus > neutral focus). When the focused word is sentence-penultimate, only the mean f_0 increases significantly across the focus conditions. See Figure 8 in Appendix C for the prediction plots.

Table 2: On-focus region: mean values of various measurements under the effect of focus, together with the results of Linear Mixed Models. *P* values smaller than 0.05 are in boldface.

	Focus condition		
	Sentence-initial focus		
Measurements	**Neutral-focus**	**Information-focus**	**Contrastive-focus**
Mean f_0 (Hz)	M=236.64, SD=23.47	M=242.90, SD=26.67	M=249.575, SD=25.62
	χ^2=5.86, df=2, p=0.053		
Max f_0 (Hz)	M=257.91, SD=32.22	M=268.10, SD=38.43	M=277.65, SD=36.73
	χ^2=5.84, df = 2, p= 0.054		
Min f_0 (Hz)	M=208.38, SD=13.66	M=207.64 , SD=12.62	M= 207.05, SD=13.64
	χ^2=0.86, df=2, p= 0.65		
Excursion size (st.)	M=3.620, SD=1.647	M=4.223 , SD= 1.72	M=4.93, SD=1.54
	χ^2= 8.03, df=2, $\boldsymbol{p}$**=0.02**		
Duration (ms.)	M=217.77, SD=17.19	M=230.42 , SD=22.36	M 238.691, SD=24.31
	χ^2=11.26, df =2, $\boldsymbol{p}$**=0.003**		
Mean Intensity (dB)	M=69.85, SD=13.20	M=69.92 , SD=13.32	M=70.463, SD=13.12
	χ^2=7.27, df=2, $\boldsymbol{p}$**=0.03**		
	Sentence-penultimate focus		
Mean f_0 (Hz)	M=223.87, SD= 22.13	M=228.85, SD=22.70	M=234.57, SD=24.60
	χ^2=7.70, df=2, $\boldsymbol{p}$**=0.02**		
Max f_0 (Hz)	M= 242.93, SD=27.89	M=248.51, SD=28.89	M=257.07, SD=32.20
	χ^2=9.38, df=2, $\boldsymbol{p}$**=0.01**		

[11] The item level random structure contains a random intercept for sentence type, by-sentence type and speaker-by-sentence type random slopes for the main effects.

[12] Since the position and the syllable weight of the stressed syllables vary by sentence type, we included sentence type as random effects.

Min f_0 (Hz)	M=202.61, SD=16.17	M=205.05, SD=17.32	M= 205.40, SD=17.12
	χ^2=1.25, df=2, p=0.53		
Excursion size (st.)	M=3.08, SD=1.39	M=3.28, SD=1.40	M=3.82, SD=1.41
	χ^2=9.83, df =2, **p=0.01**		
Duration (ms.)	M=233.574 , SD=19.94	M=248.13 , SD=25.05	M=257.41, SD=24.58
	χ^2=10.52, df=2, **p=0.005**		
Mean Intensity (dB)	M= 67.97, SD=13.76	M= 68.51, SD=13.58	M= 69.30, SD=13.77
	χ^2=10.48, df=2, **p=0.005**		
Sentence-final focus			
Mean f_0 (Hz)	M=198.36, SD=21.81	M=202.23, SD=20.84	M=207.38, SD=21.86
	χ^2=5.46, df=2, p=0.07		
Max f_0 (Hz)	M=207.87, SD=25.31	M=210.97, SD=22.62	M=217.89, SD=25.43
	χ^2=5.97, df=2, p=0.05		
Min f_0 (Hz)	M=185.92, SD=17.68	M=188.29, SD=18.57	M=188.59, SD=18.16
	χ^2=1.106, df=2, p=0.57		
Excursion size (st.)	M=1.80, SD=0.89	M=1.926, SD=0.84	M=2.46, SD=1.075
	χ^2=8.28, df=2, **p=0.01**		
Duration (ms.)	M=173.30, SD=27.75	M=179.75, SD=34.890	M=190.62, SD=31.27
	χ^2=6.23, df=2, **p=0.04**		
Mean Intensity (dB)	M=64.28, SD=12.984	M= 65.27, SD=12.92	M= 66.14, SD=13.25
	χ^2=8.45, df=2, **p=0.01**		

The post-hoc comparisons of the three focus conditions are displayed in Table 3. The table shows the following results. First, excursion size is more expanded in contrastive focus than in both neutral focus and information focus. Second, duration is longer in contrastive focus than in neutral focus. In addition, duration is longer in information focus than in neutral focus, but only when the focused word is sentence-penultimate. Third, mean intensity is stronger in contrastive focus than in neutral focus. Moreover, mean intensity is higher in contrastive focus than in information focus when the focus word is sentence-initial and -penultimate. When the focused word is sentence-final, mean intensity is higher in information focus than in neutral-focus. Fourth, maximum f_0 is higher in contrastive focus than in neutral-focus, and higher than in information-focus when the focused word is sentence-penultimate and -final. Finally, when the focused word is sentence-penultimate, mean f_0 is higher in contrastive focus than in both neutral-focus and information-focus.

Table 3: Post–hoc comparisons of effects of focus on the stressed syllable of on-focused words, after Tukey adjustments. *P* values smaller than 0.05 are in boldface. *Sentential position* refers to the sentential position of the focused word in the target sentence.

		Focus Condition		
Measurements	**Sentential Position**	**Neutral Focus vs. Information Focus**	**Neutral Focus vs. Contrastive Focus**	**Information Focus vs. Contrastive Focus**
Excursion size (st.)	initial	p = 0.20	**p = 0.002**	**p = 0.002**
	penultimate	p = 0.41	**p = 0.0001**	**p = 0.01**
	final	p = 0.62	**p = 0.0002**	**p = 0.001**
Duration (ms.)	initial	p = 0.052	**p = 0.001**	p = 0.22
	penultimate	**p = 0.0001**	**p = 0.0004**	p = 0.09
	final	p = 0.56	**p = 0.006**	p = 0.09
Mean Intensity (dB)	initial	p = 0.93	**p = 0.03**	**p = 0.02**
	penultimate	p = 0.22	**p = 0.0001**	**p = 0.01**
	final	**p = 0.003**	**p = 0.005**	p = 0.12
Max f_0 (Hz)	penultimate	p = 0.228	**p = 0.001**	**p = 0.02**
Mean f_0 (Hz)	penultimate	p = 0.25	**p = 0.01**	**p = 0.02**

In short, contrastive focus is acoustically marked with an expanded excursion size, increased duration, increased intensity, higher maximum f_0, and higher mean f_0, compared to other elements that are under information focus or neutral focus. This confirms our visual observation of Figure 3 with regard to the on-focus region.

B. Post-focus Region

Table 4 presents the means and standard deviations of the acoustic measurements (13) of the stressed syllables of the post-focus words. It shows that the mean values of all acoustic measurements, apart from excursion size and duration in sentence-penultimate focus, significantly decreases across the focus conditions: neutral-focus < information-focus < contrastive-focus. See Figure 9 in Appendix C for the prediction plots.

Table 4: Post-focus region: mean values of various measurements under the effect of focus, together with results of Linear Mixed Models. P values smaller than 0.05 are in boldface.

Focus condition			
Sentence-initial focus			
Measurements	Neutral-focus	Information-focus	Contrastive-focus
Mean f_0 (Hz)	M=215.65 , SD=17.18	M=205.90 , SD=15.12	M=205.74, SD=18.88
	χ^2=14.002, df =2, **p=0.001**		
Max f_0 (Hz)	M=229.28 , SD=21.90	M=219.47 , SD=18.71	M=219.96, SD=21.65
	χ^2=9.67, df=2, **p=0.01**		
Min f_0 (Hz)	M=204.01 , SD=14.25	M=194.41 , SD=13.13	M=194.07, SD=17.31
	χ^2=13.41, df=2, **p=0.001**		
Excursion size (st.)	M=1.89 , SD=0.85	M=1.977 , SD=0.71	M=2.03, SD=0.66
	χ^2=1.78, df = 2, p=0.41		
Duration (ms.)	M=145.10 , SD=9.02	M=139.36 , SD=9.59	M= 138.54, SD=10.18
	χ^2=6.49, df=2, **p=0.04**		
Mean Intensity (dB)	M=64.61 , SD=12.91	M=63.17 , SD=13.35	M=62.94, SD=13.02
	χ^2=13.40, df=2, **p=0.001**		
Sentence-penultimate focus			
Mean f_0 (Hz)	M=194.60, SD=20.61	M=188.24, SD= 18.80	M=181.24, SD=24.03
	χ^2=7.84, df=2, **p=0.02**		
Max f_0 (Hz)	M=200.53, SD=21.17	M=193.44, SD=18.06	M=188.47, SD=21.94
	χ^2=7.84, df=2, **p=0.02**		
Min f_0 (Hz)	M=184.37, SD= 21.79	M=178.83, SD= 25.68	M=170.03, SD=29.37
	χ^2=9.66, df=2, **p=0.01**		
Excursion size (st.)	M=1.540, SD=1.37	M=1.523, SD=1.87	M= 2.04, SD=1.93
	χ^2=4.31, df=2, p=0.12		
Duration (ms.)	M=142.73, SD=11.41	M=143.44, SD=12.56	M=145.60, SD=12.63
	χ^2=2.61, df=2, p=0.30		
Mean Intensity (dB)	M=63.79, SD=13.37	M=62.79, SD=13.49	M=62.68, SD=13.53
	χ^2=7.07, df=2, **p=0.03**		

Table 5 displays post-hoc comparisons of the three focus conditions: neutral-focus, information-focus, and contrastive-focus. Significantly, maximum f_0 and mean intensity of post-focus words are more compressed in information focus and contrastive focus than in their neutral-focus counterparts, irrespective of the sentential position of the focused word. Table 5 also shows that when the focused word is sentence-initial, mean f_0, minimum f_0 and duration of the stressed syllable of the post-focus words are more compressed in information focus and contrastive focus conditions than their neutral-focus counterparts. Also, when the focused word is sentence-penultimate, mean f_0 and minimum f_0 of the stressed syllables of the post-focus words are more compressed in contrastive focus than in both information focus and neutral focus conditions.

Table 5: Post–hoc comparisons of effects of focus on the stressed syllable of post-focused words, after Tukey adjustments. P values smaller than 0.05 are in boldface. *Sentential position* refers to the sentential position of the focused word in the target sentence.

		Focus Condition		
Measurements	**Sentential Position**	**Neutral Focus vs. Information Focus**	**Neutral Focus vs. Contrastive Focus**	**Information Focus vs. Contrastive Focus**
Mean f_0 (Hz)	initial	**$p = 0.0001$**	**$p = 0.0003$**	$p = 1.0$
	penultimate	$p = 0.053$	**$p = 0.002$**	**$p = 0.03$**
Max f_0 (Hz)	initial	**$p = 0.001$**	**$p =0.002$**	$p = 0.96$
	penultimate	**$p = 0.02$**	**$p = 0.01$**	$p = 0.12$
Min f_0 (Hz)	initial	**$p = 0.0001$**	**$p = 0.0001$**	$p = 0.98$
	penultimate	$p = 0.15$	**$p = 0.0006$**	**$p = 0.003$**
Mean Intensity (dB)	initial	**$p = 0.0001$**	**$p = 0.0001$**	$p = 0.80$
	penultimate	**$p = 0.002$**	**$p = 0.0003$**	$p = 0.87$
Duration (ms.)	initial	**$p = 0.005$**	**$p = 0.02$**	$p = 0.87$

In summary, there is evidence of post-focus compression of both maximum f_0 and mean intensity. This indicates that post-focus compression is active in EA, which supports Blodgett et al.'s (2007) preliminary visual and auditory investigation of EA intonation mentioned earlier in §1.3. As for the difference between information focus and contrastive focus, they show significant differences only in mean f_0 and min f_0.

C. Pre-focus Region

Table 6 displays the means and standard deviations of the acoustic measurements of stressed syllables of pre-focus words across the focus conditions: neutral focus, information focus and contrastive focus. As is apparent, there is a significant difference across the focus conditions in excursion size and duration of the stressed syllables of pre-focus words. See Figure 10 in Appendix C for the prediction plots.

Table 6: Pre-focus region: mean values of various measurements under the effect of focus, together with results of Linear Mixed Models. P values smaller than 0.05 are in boldface.

	Focus condition		
	Sentence-penultimate focus		
Measurements	Neutral-focus	Information-focus	Contrastive-focus
Mean f_0 (Hz)	M=226.24, SD=19.27	M=222.81, SD=17.164	M=224.79, SD= 17.82
	χ^2=3.31, df=2, p=0.19		
Max f_0 (Hz)	M=243.35, SD=25.12	M=234.31, SD=21.99	M=235.63, SD=22.25
	χ^2=4.92, df=2, p=0.08		
Min f_0 (Hz)	M=203.47, SD=14.29	M=205.37, SD=13.63	M=207.41, SD=13.20
	χ^2=3.94, df=2, p=0.14		
Excursion size (st.)	M=3.02, SD=1.06	M= 2.29, SD=0.96	M=2.21, SD=1.03
	χ^2=3.85, df=2, p=0.14		
Duration (ms.)	M=196.44, SD=17.74	M=179.50, SD= 18.88	M=178.61, SD=21.33
	χ^2=11.92, df=2, **p=0.002**		
Mean Intensity (dB)	M=69.25, SD=13.64	M=69.18, SD=13.53	M=69.35, SD=13.42
	χ^2=0.56, df=2, p=0.75		
	Sentence-final focus		
Mean f_0 (Hz)	M=224.78, SD=17.99	M=223.04, SD=18.10	M=224.591, SD=19.63
	χ^2=2.62, df=2, p=0.27		
Max f_0 (Hz)	M=237.72, SD=22.90	M=232.67, SD=21.13	M=234.37, SD=22.13
	χ^2 = 5.12, df = 2, p=0.08		
Min f_0 (Hz)	M=208.37, SD=14.23	M=209.75, SD=14.46	M= 210.59, SD=15.51
	χ^2=2.61, df=2, p=0.27		

Excursion size (st.)	M=2.23, SD=1.07	M=1.71, SD=0.74	M=1.814, SD=0.66
	χ^2=10.31, df=2, ***p*=0.01**		
Duration (ms.)	M=163.89, SD=11.70	M=153.39, SD=11.59	M=155.44, SD=13.15
	χ^2=13.80, df = 2, ***p*=0.001**		
Mean Intensity (dB)	M=68.82, SD=13.08	M=68.66, SD=13.10	M=68.77, SD=13.29
	χ^2=0.77, df=2, *p*=0.68		

The post-hoc tests in Table 7 show that duration of stressed syllables of pre-focus words preceding information focus and contrastive focus is significantly shorter than that of their neutral-focus counterparts for both Sentence-penultimate and sentence-final focus. But Excursion size of stressed syllables of pre-focus words preceding information focus and contrastive focus, however, is significantly more compressed than that of their neutral-focus counterparts only for sentence-final focus. Meanwhile, the difference between information focus and contrastive focus in the pre-focus region is not statistically significant.

Table 7: Post–hoc comparisons of effects of focus on the stressed syllable of pre-focused words, after Tukey adjustments. *P* values smaller than 0.05 are in boldface. *Sentential position* refers to the position of the focused word in the target sentence.

		Focus Condition		
Measurements	**Sentential Position**	**Neutral Focus vs. Information Focus**	**Neutral Focus vs. Contrastive Focus**	**Information Focus vs. Contrastive Focus**
Excursion size (st.)	final	$\boldsymbol{p = 0.0004}$	$\boldsymbol{p = 0.03}$	$p = 0.70$
Duration (ms.)	penultimate	$\boldsymbol{p = 0.001}$	$\boldsymbol{p = 0.001}$	$p = 0.92$
	final	$\boldsymbol{p = 0.001}$	$\boldsymbol{p = 0.001}$	$p = 0.40$

In summary, there are acoustic differences in pre-focus region across the focus conditions, related to excursion size and duration. Duration of pre-focus region under information focus and contrastive focus is shorter than that under neutral focus. In addition, excursion size of pre-focus region under sentence-final information focus and contrastive focus is less expanded than that under neutral focus. This indicates that there is pre-focus compression of duration and excursion size in EA, similar to that reported for Lebanese Arabic (Chahal, 2001). However, this pre-focus compression, interestingly, has not been observed in other dialects, such as Egyptian Arabic (Hellmuth, 2006*b*), Hijazi Arabic (Alzaidi et al., 2019) or Makkan Arabic (Alzaidi, 2022).

The data (i.e., measured values used for statistical analyses) and the reproducible analysis scripts are publicly available at Open Science Framework (OSF) (`https://osf.io/w5qvh`).

2.2. Discussion

The above results demonstrate that there is a specific intonational pattern associated with focus in EA. This prosodic pattern shares some features with the patterns found in other Arabic dialects reviewed so far such as Egyptian, Hijazi, and Lebanese Arabic. However, prosodic differences exist across these dialects, in particular Hijazi Arabic, which is also a Gulf Arabic dialect similar to EA.

In the post-focus region, when focus is either sentence-initial or sentence-penultimate, maximum f_0 and mean intensity of stressed syllables of post-focus words show systematic differences across the three focus conditions. That is, when focus is sentence-initial or sentence-penultimate, maximum f_0 of post-focused words in information focus and contrastive focus is lower than that of their neutral-focus counterparts in the same structure. In addition, mean intensity of post-focused words in information focus and contrastive focus is weaker than that of their neutral-focus counterparts. When focus is sentence-initial, mean f_0 and minimum f_0 of post-focus words in information focus and contrastive focus are lower than those of their neutral-focus counterparts. When focus is sentence-penultimate, mean f_0 and minimum f_0 of post-focus words in contrastive focus are lower than those of their neutral-focus counterparts. In terms of duration, when focus is sentence-initial, post-focus words in information focus and contrastive focus are shorter than those of their neutral-focus counterparts. These results show that post-focus compression is in maximum f_0 and mean intensity when focus is sentence-initial and sentence-penultimate, and in mean f_0, minimum f_0 and duration only when focus is sentence-initial. Given these results, we conclude that EA is a +PFC language, alongside the other Arabic dialects studied so far, including Egyptian, Hijazi and Lebanese Arabic (reviewed in §1.1.2).

In the on-focus region, max f_0, min f_0, excursion size, duration and mean intensity of contrastive focus (but not information focus) are all higher than those of their counterparts under neutral focus but only duration and mean intensity of information focus have greater values than those in neutral focus.

With regard to the pre-focus region, the acoustic analyses have shown that when focus is either sentence-penultimate or sentence-final, excursion size and duration of pre-focus region are lower than in their neutral-focus counterparts.

The acoustic analyses, therefore, provided answers to the research questions raised in (10a and b), repeated below for convenience with the answers obtained from the production experiment.

(14) a. Does focus in Emirati Arabic involve tri–zone prosodic adjustments as found in many other languages?

> Yes, to a large extent.

b. Are there prosodic differences between information focus and contrastive focus?

> Yes, with exceptions.

The present results are consistent with the preliminary finding of Blodgett et al. (2007) that there is post-focus compression due to a discourse function. The current study empirically confirms that focus is the discourse function that involves post-focus compression. The results of the current study are also consistent with the findings of Chahal (2001), Hellmuth (2006*b*), Hellmuth (2009) and Alzaidi et al. (2019) for other Arabic dialects, including Egyptian, Hijazi and Lebanese Arabic. However, post-focus compression in EA is in maximum f_0 and mean intensity but not in excursion size (i.e., pitch range), as is also the case in Egyptian (Hellmuth, 2006*b*) and Hijazi Arabic (Alzaidi et al., 2019). Within Xu's (2011) distribution of languages, the present results place EA as a +PFC language alongside the other Arabic dialects studied so far, excluding Makkan Arabic which is -PFC (Alzaidi, 2022).

Another finding is that there are systematic differences between contrastive focus and neutral focus. The contrastive-focus words had more expanded excursion size, stronger intensity, and longer duration than their neutral-focus counterparts in all sentential positions. The difference between contrastive-focus and information-focus words is only in excursion size in all sentential positions, and in mean intensity only in sentence-initial and sentence-penultimate positions. These results with regard to contrastive focus are similar to what has been found in Egyptian Arabic (Hellmuth, 2006*b*, 2009) as well as Hijazi Arabic (Alzaidi et al., 2019), in that excursion size of contrastive focus is more expanded than in neutral focus and information focus.

3. PENTAtrainer Modelling

The previous section has shown that focus intonation in EA is characterized by an expansion of excursion size of on-focus word (in particular, contrastive focus), a reduction of maximum f_0 of post-focus words, and a reduction of excursion size of pre-focus words. In this section, we test whether these focus intonation patterns can be captured in computational modelling based on PENTA, an articulatory-functional model. The modelling will also examine what other factors besides focus may contribute to the fully detailed f_0 contours in EA. The quality of computational modelling is assessed by three criteria: numerical synthesis accuracy (§3.2.2), visual comparison between the synthetic f_0 contours and the original, and perceptual appraisal. The following are the research questions repeated from (10c, d) for convenience.

(15) a. Can Emirati Arabic focus intonation be computationally captured by an articulatory-functional model?

b. What can computational modelling tell us about Emirati Arabic intonation that we cannot easily learn from acoustic analysis alone?

3.1. Methods

Following previous modelling practices (Raidt et al., 2004; Sakurai et al., 2003; Sun and Xu, 2002; Xu and Prom-on, 2014), we first separated the EA data from the production experiment (§3.3) into a learning subset and a testing subset. The learning subset was used to train PENTAtrainer to extract target parameters (§3.1.1). The testing subset was used to examine how well the learned parameters could predict EA intonation. The learning subset consisted of 1620 utterances by nine native speakers (9 speakers x 4 words x 3 foci x 3 sets of sentences x 5 repetitions = 1620). The testing subset consisted of 360 utterances from two speakers (2 speakers x 4 words x 3 foci x 3 sets of sentences x 5 repetitions = 360) selected from the total of 11 speakers (§2.1.2) using the mean f_0 across all repetitions of each subject as an arbitrary criterion. One speaker (coded F7) had the highest mean f_0 across all repetitions, and the second speaker (coded F5) had the lowest mean f_0 across all repetitions. The criterion was to make sure that the selection was arbitrary rather than based on any subjective choice that may carry researcher bias.

3.1.1. Modelling

As introduced in §1.2.2, PENTAtrainer learns a set of target parameters that simultaneously represent a number of communicative functions. The experimental data in the production study were systematically controlled for focus, but the other conditions (i.e. Stress/Sposition/Weight/Pword as illustrated in 16 below) were not systematically controlled because their functions and structures, in particular, whether they are all independent contributors to surface f_0 contours, are unclear. Therefore, our modelling objective was to both capture focus intonation of EA, and explore the other functions that are still putative.

The acoustic analysis of the production data has established that in EA, the raising of peak f_0 and expansion of excursion size under focus and the f_0 lowering and compression of excursion size are most clearly manifested in stressed syllables as mentioned in §2.1.5. It is therefore key to modelling to specify not only the division of the focus regions, but also the location of the stressed syllables, as they are the sites where the focus effects are manifested the most.

As reviewed in §1.3, stress is jointly predictable in Emirati Arabic by syllable weight and syllable position in word. Thus, for a syllable to be stressed in EA, we need to first specify the temporal scope of Pword, then syllable weight, and finally word-level stressed syllable position. We therefore decided to test whether it is beneficial to specify **Focus**, **Stress**, **Position of word-level stress, Weight, and Pword (= prosodic word), as listed in (16)**. In (16) and hereafter, the names of the annotated functions are in boldface, while the function-internal categories are in italics.[13]

(16) a. **Focus**: pre-focus (*PRE*), on-focus (*ON*) and post-focus region (*POS*)

b. **Stress** : unstressed (*U*), stressed (*S*)

c. **Sposition** – Position of word-level stress: initial (*I*), penultimate (*P*), final (*F*)

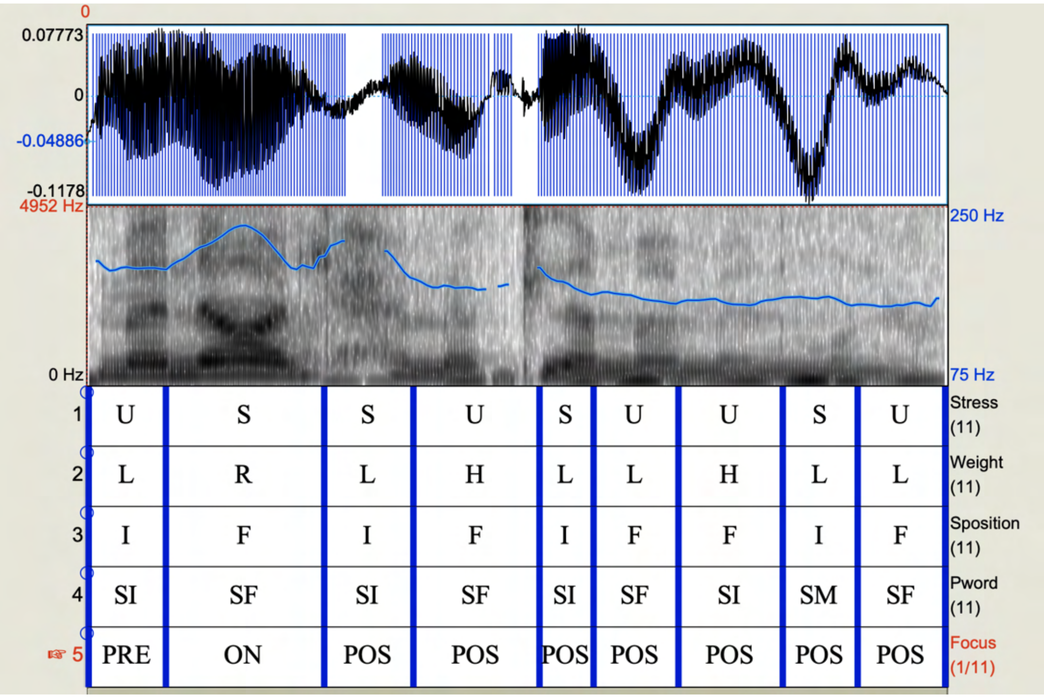
d. **Weight**: superheavy (*R* = CVVC, CVCC), heavy (*H* = VV, CVC), light (*L* = CV)

e. **Pword** – Position of the syllable in a group of words in which only one syllable sounds stressed: initial (*SI*), medial (*SM*) and final position (*SF*) (cf. Holes, 1990; Hellmuth, 2006*b*, 2007; Watson, 2002).

Figure 4 illustrates the annotation of the five functional layers: **Focus**, **Stress**, **Sposition**, **Weight** and **Pword**. Each layer was annotated independently and the function-internal categories were defined by the first author. The interval boundaries within each layer were marked according to the time span of the prosodic event, as defined by the first author. As pointed out earlier in §1.2.2, names of categories do not carry any phonetic specifications other than their identity.

The annotation in Figure 4 also exemplifies how PENTAtrainer-based modelling as shown in Figure 1 and Figure 2 works. Figure 1 shows that during articulation (the rightmost block), speakers produce only a single sequence of pitch targets, each of which is jointly determined (center block) by multiple functions (leftmost block). In PENTAtrainer, each multi-functional target is defined by concatenating the category names of all the functional layers, as illustrated in Figure 2. For example, for the utterance in Figure 4, the first multi-functional target is ULISIPRE, the second target is SRFSFON, and so on. For each of these targets, PENTAtrainer aims at finding a set of target parameters (m, b, and λ) that can generate synthetic f_0 contours which are close to all the original f_0 contours of the same target category, collectively. The search for the optimal targets is done through

[13] The list of the functions in 16 is presented in the order in which they will be presented in the results.



1	U	S	S	U	S	U	U	S	U	Stress (11)
2	L	R	L	H	L	L	H	L	L	Weight (11)
3	I	F	I	F	I	F	F	I	F	Sposition (11)
4	SI	SF	SI	SF	SI	SF	SI	SM	SF	Pword (11)
☞ 5	PRE	ON	POS	POS	POS	POS	POS	POS	POS	Focus (1/11)

Figure 4: /Ma.**nāl** **ḥa**.mat **La**.ma min **Mu**.na/ "Manal protected Lama from Muna". An example of the conversation process from the parallel functional annotation to the essential functional combinations. For "**Stress**", *S* denotes stressed syllables and *U* denotes unstressed syllables. For "**Weight**" layer, *R*, *H*, *L* denote superheavy, heavy and light syllables, respectively. For "**Sposition**" (i.e., Position of word-level stress), *I*, *P* and *F* denote initial, penultimate and final position, respectively. For "**Pword**" (i.e., the position in the prosodic word), *SI*, *SM*, *SF* denote initial, medial and final position, respectively. For "**Focus**" layer, *PRE*, *ON*, *POST* denote pre-focus, on-focus and post-focus region, respectively.

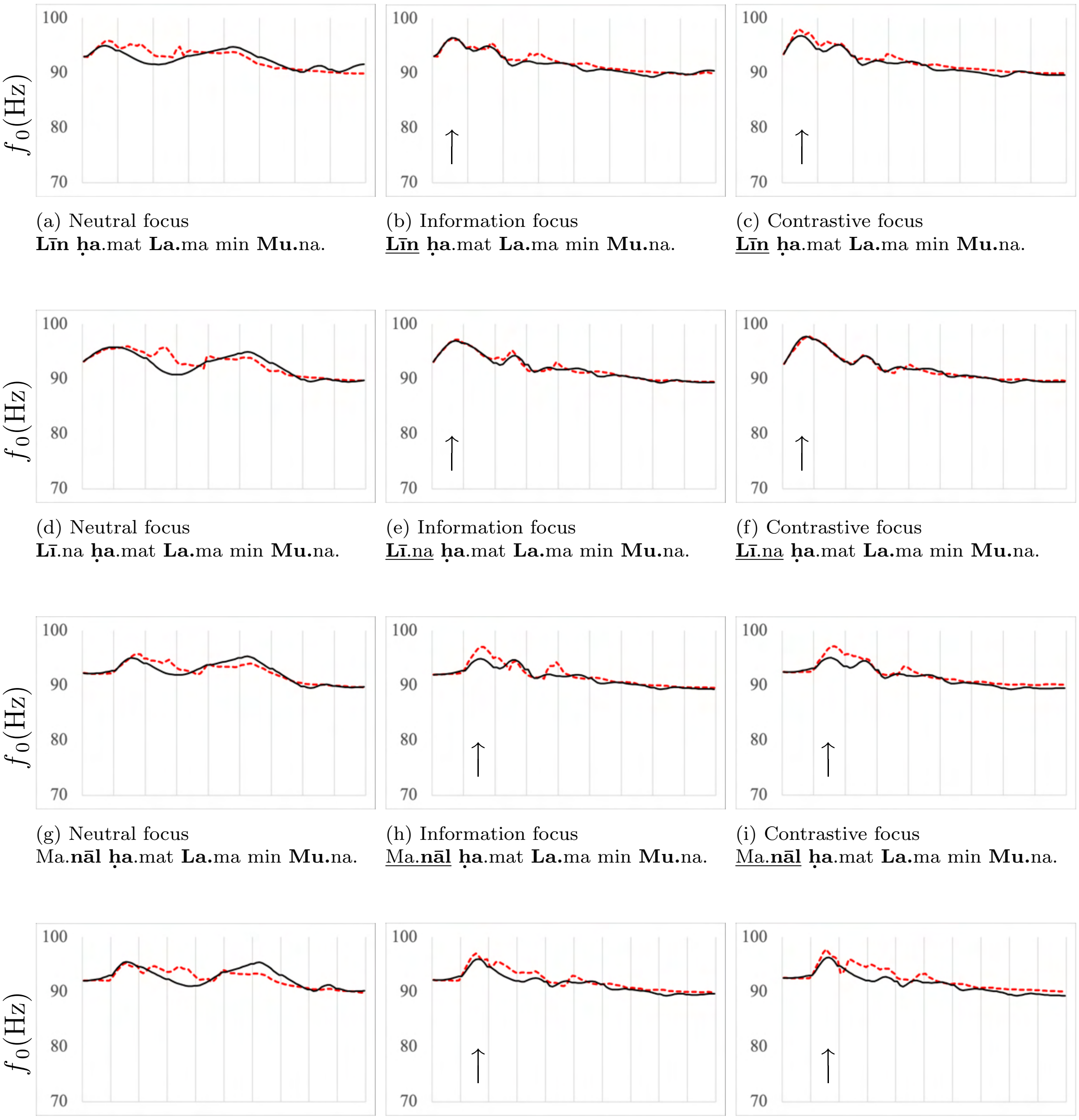
analysis-by-synthesis based on simulated annealing (Kirkpatrick et al., 1983). In each search cycle, a whole set of parameters for all the multi-functional targets is randomly set and then used to synthesize f_0 contours of all the utterances through qTA. The parameter set that happens to result in better overall f_0 fit (as measured by total sum of square error between original and synthesized f_0 contours) is adopted if it is also above the acceptance probability determined by the annealing temperature at the current learning stage. The whole set of multi-functional targets obtained at the end of the learning period (750 cycles) is adopted as the learned targets, which are then used for testing and evaluation.

In order to compare the contribution of each of the five functions (16), learning and synthesis were performed using the learning set first with all five functional layers included, and then with each of the functional layers removed. With all the 5 functions included, there were 26 different combinations of communicative functions to be learned from the dataset. Excluding each of the 5 functions results in different reductions of the total number of combinations of parameters (see Appendix D for the set of parameters when each function is removed). The parameters are automatically optimized by the Learn tool in PENTAtrainer. The default optimization parameters were: Maximum Iteration = 700, Learning Rate = 0.1, Starting Temperature: =700, Reduction Factor = 0.97, Silent Threshold = 0.2.

3.2. Numerical results

3.2.1. Accuracy of model predictions

Figure 5 shows graphical comparisons between the synthesized f_0 and the natural f_0 of the two speakers in the testing subset. The synthesized f_0 contours in each figure were generated from the 26 multi-function parameters in Table 8. These parameters were extracted from the nine speakers in the learning subset (§3.1). Both the natural and synthesized f_0 contours were averaged across the 10 repetitions by the two speakers for each focus condition of each target sentence. The f_0 contours are time-normalised with regard to the syllable.



(a) Neutral focus
Līn **ḥa**.mat **La.**ma min **Mu.**na.

(b) Information focus
Līn **ḥa**.mat **La.**ma min **Mu.**na.

(c) Contrastive focus
Līn **ḥa**.mat **La.**ma min **Mu.**na.

(d) Neutral focus
Lī.na **ḥa**.mat **La.**ma min **Mu.**na.

(e) Information focus
Lī.na **ḥa**.mat **La.**ma min **Mu.**na.

(f) Contrastive focus
Lī.na **ḥa**.mat **La.**ma min **Mu.**na.

(g) Neutral focus
Ma.**nāl** **ḥa**.mat **La.**ma min **Mu.**na.

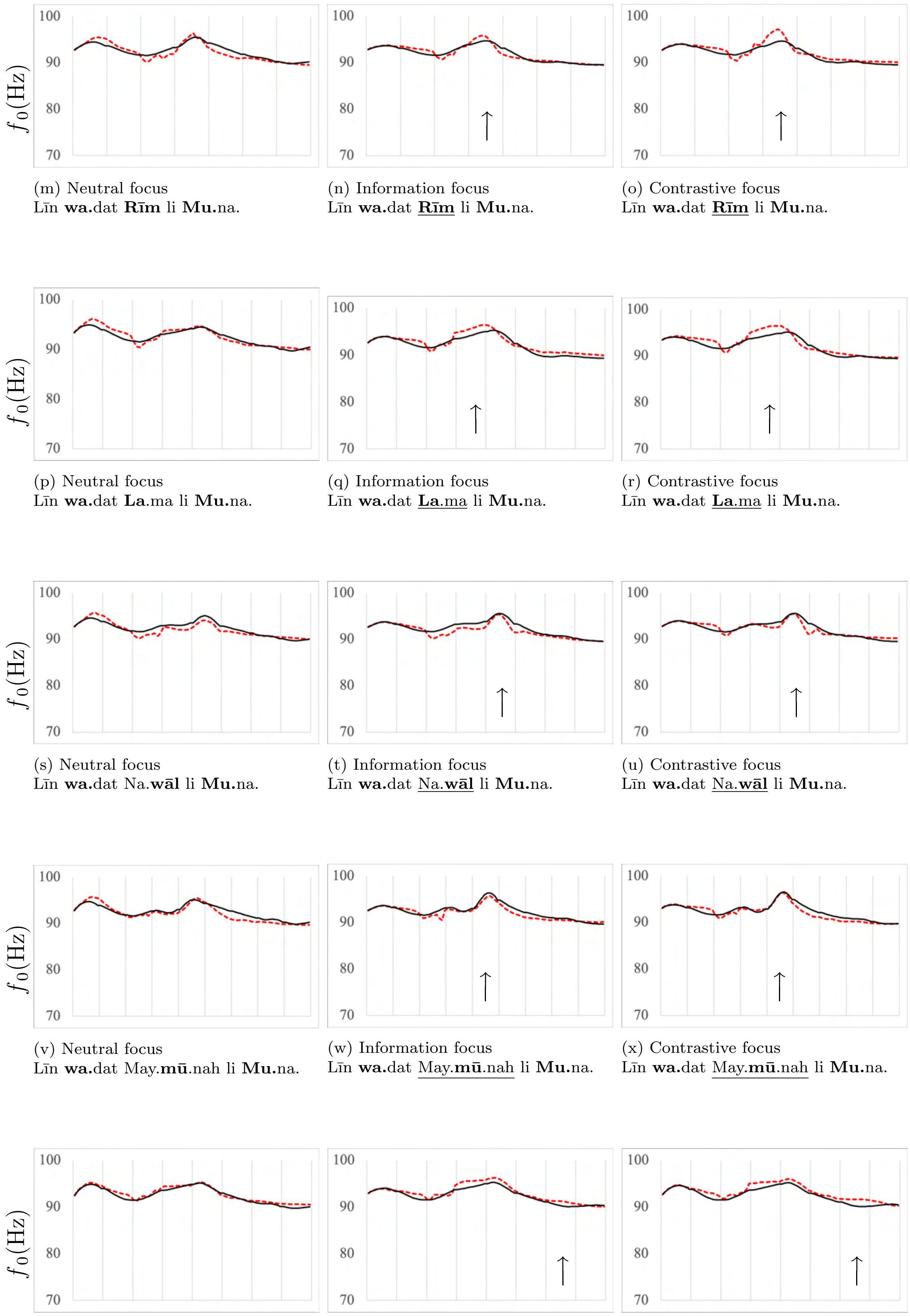
(h) Information focus
Ma.**nāl** **ḥa**.mat **La.**ma min **Mu.**na.

(i) Contrastive focus
Ma.**nāl** **ḥa**.mat **La.**ma min **Mu.**na.

(j) Neutral focus
Ma.**lī**.kah **ḥa**.mat **La.**ma min **Mu.**na.

(k) Information focus
Ma.**lī**.kah **ḥa**.mat **La.**ma min **Mu.**na.

(l) Contrastive focus
Ma.**lī**.kah **ḥa**.mat **La.**ma min **Mu.**na.



(m) Neutral focus
Līn **wa.**dat **Rīm** li **Mu.**na.

(n) Information focus
Līn **wa.**dat **Rīm** li **Mu.**na.

(o) Contrastive focus
Līn **wa.**dat **Rīm** li **Mu.**na.

(p) Neutral focus
Līn **wa.**dat **La.**ma li **Mu.**na.

(q) Information focus
Līn **wa.**dat **La.**ma li **Mu.**na.

(r) Contrastive focus
Līn **wa.**dat **La.**ma li **Mu.**na.

(s) Neutral focus
Līn **wa.**dat Na.**wāl** li **Mu.**na.

(t) Information focus
Līn **wa.**dat Na.**wāl** li **Mu.**na.

(u) Contrastive focus
Līn **wa.**dat Na.**wāl** li **Mu.**na.

(v) Neutral focus
Līn **wa.**dat May.**mū**.nah li **Mu.**na.

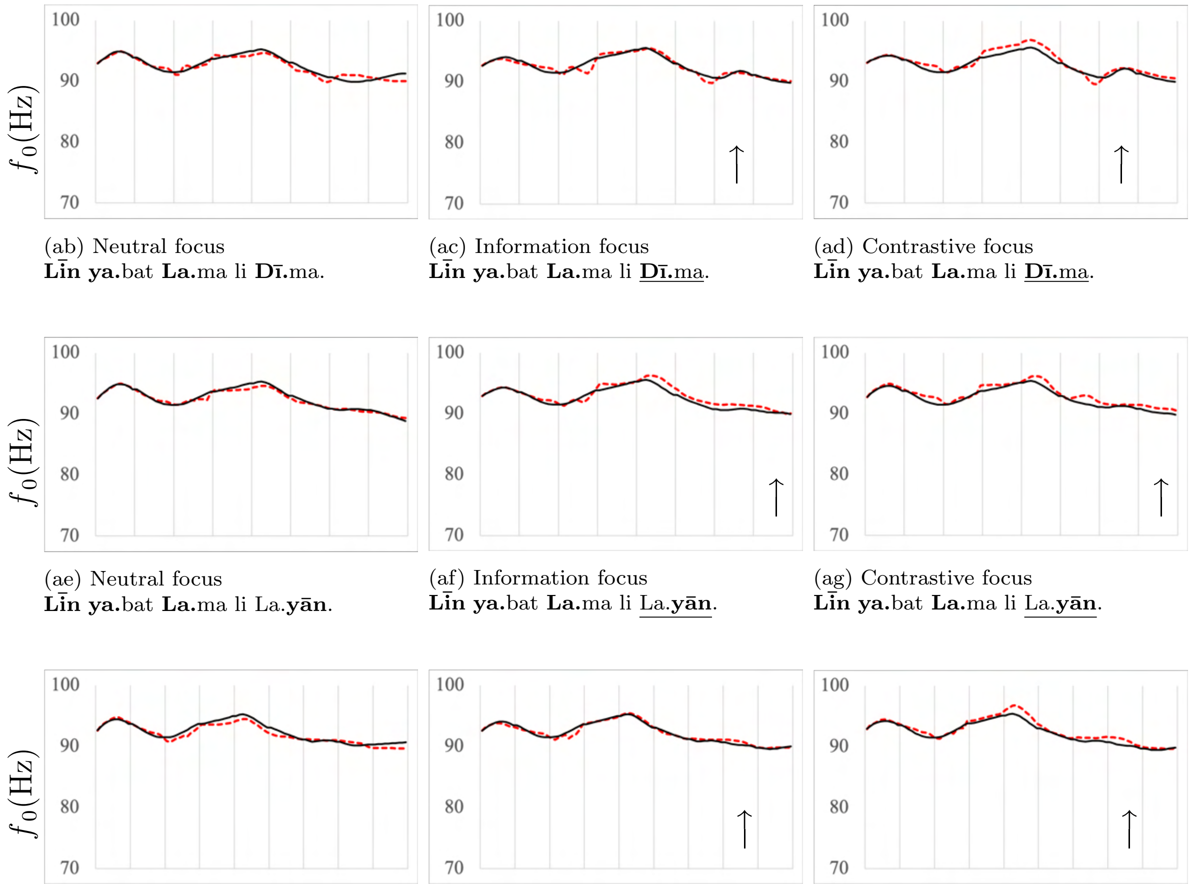
(w) Information focus
Līn **wa.**dat May.**mū**.nah li **Mu.**na.

(x) Contrastive focus
Līn **wa.**dat May.**mū**.nah li **Mu.**na.

(y) Neutral focus
Līn ya.bat **La.**ma li **Mu.**na.

(z) Information focus
Līn ya.bat **La.**ma li **Mu.**na.

(aa) Contrastive focus
Līn ya.bat **La.**ma li **Mu.**na.



(ab) Neutral focus
Līn **ya.**bat **La.**ma li **Dī.**ma.

(ac) Information focus
Līn **ya.**bat **La.**ma li **Dī.**ma.

(ad) Contrastive focus
Līn **ya.**bat **La.**ma li **Dī.**ma.

(ae) Neutral focus
Līn **ya.**bat **La.**ma li La.**yān**.

(af) Information focus
Līn **ya.**bat **La.**ma li La.**yān**.

(ag) Contrastive focus
Līn **ya.**bat **La.**ma li La.**yān**.

(ah) Neutral focus
Līn **ya.**bat **La.**ma li Mu.**nī.**rah.

(ai) Information focus
Līn **ya.**bat **La.**ma li Mu.**nī.**rah.

(aj) Contrastive focus
Līn **ya.**bat **La.**ma li Mu.**nī.**rah.

Figure 5: Mean time-normalised natural (red dotted line) and synthetic (black solid line) f_0 contours averaged across 10 repetitions of the two speakers not used in the model training. The Y-axis displays f_0 values in semitone. The vertical lines mark syllable boundaries. Underline indicates a focus placement and bold-face indicates a stressed syllable of that word. All the synthetic contours were generated with parameters (**Stress**: *S* and *U*, **Weight**: *R*, *H*, *L*, **Sposition**: *I*, *P*, *F*, **Pword**: *SI*, *SM*, *SF*, **Focus**: *Pre*, *On*, *Pos*). The upper arrow indicates the position of the stressed syllable of on-focus words.

The f_0 plots in Figure 5, clearly show the accuracy of the model's prediction of the natural f_0 of each of the focus conditions by the two speakers not included in the model training. First, the synthesized f_0 of the word under focus in Figure 5b, 5e, 5f, 5ac, 5ad and Figure 5ai is similar to the natural f_0. The f_0 of /Līna/ is closely predicted by the model when it is information-focus as in Figure 5e and also when it is contrastive-focus as in Figure 5f. In addition, the f_0 of the target word /Dīma/ is predicted by the model when it is information-focus as in Figure 5ac and also when it is contrastive-focus as in Figure 5ad. The model predicted the f_0 of the focused word /Līn/ when it is information-focus (Figure 5b) but not when it is contrastive-focus (Figure 5c). In addition, the model predicted the f_0 of focused word /Munīrah/ when it is information-focus (Figure 5ai) but not when it is contrastive-focus (Figure 5aj). When /Līn/ and /Munīrah/ are under contrastive-focus, their synthesized f_0 is lower than the natural f_0. This is noticeable not only when the target word is contrastive-focus, but also when the target word is information-focus. In Figure 5h, 5i, 5k, 5l, 5n, 5o, 5q, 5r, 5z, 5aa, 5af and Figure 5ag, the synthesized f_0 of the focused word is lower than the natural f_0. Furthermore, there are cases in which the synthesized f_0 of the focused word is higher than the natural f_0 as in Figure 5t, 5u, 5w and Figure 5x. These cases are observed only when the focused word is sentence-penultimate. Interestingly, when /Nawāl/ and /Maymūnah/ are contrastive-focus as in Figure 5u and 5x, respectively, their synthesized f_0 does not differ much from the natural f_0 as in Figure 5t and 5w when they are information-focus.

Table 8: The learned parameter values from the 9 speakers. For **stress** function, *S* denotes stressed syllables and *U* denotes unstressed syllables. For **weight**, *R*, *H*, *L* denote superheavy, heavy, and light syllable, respectively. For **Sposition**, *I*, *P*, and *F* denote initial, penultimate, and final position, respectively. For **Pword**, *SI*, *SM*, *SF* denote initial, medial, and final position, respectively. For **focus**, *PRE*, *ON*, *POST* denote pre-focus, on-focus, and post-focus regions, respectively.

Stress	**Weight**	**Sposition**	**Pword**	**Focus**	$\boldsymbol{m}$ (st/s)	$\boldsymbol{b}$ (st)	λ
S	H	I	SI	PRE	-10.02	4.74	16.52
S	H	I	SM	PRE	-72.07	6.82	7.62
S	H	I	SI	ON	-20.7	5.6	18.35
S	H	I	SM	ON	-45.43	0.03	27.79
S	H	P	SM	ON	-52.98	2.76	16.28
S	H	P	SM	PRE	-40.88	4.77	10.05
S	L	I	SI	PRE	-5.4	1.01	8.74
S	L	I	SM	PRE	-21.75	-1.92	15.4
S	L	I	SI	POS	-21.65	-0.74	68.65
S	L	I	SM	POS	-15.88	-2.33	22.51
S	L	I	SM	ON	-51.96	0.71	31.32
S	L	I	SI	ON	-6.55	2.15	50.84
S	R	F	SI	ON	-52.83	1.72	20.09
S	R	F	SF	ON	-33.27	0.67	13.82
S	R	F	SI	PRE	-35.08	3.65	15.94
S	R	F	SF	PRE	-17.02	0.33	10.07
U	L	F	SF	PRE	-16.97	1.54	13.1
U	L	F	SI	PRE	2.76	-1.1	40.53
U	L	F	SI	POS	-6.78	-2.05	100
U	L	F	SF	POS	-35.5	-1.06	27.54
U	L	I	SI	PRE	44.4	-2.09	24.62
U	L	I	SM	PRE	-14.31	-10.36	4.71
U	H	F	SF	PRE	8.45	0.09	15.43
U	H	F	SI	PRE	-0.65	-1.39	35.27
U	H	F	SF	POS	0.8	-0.44	20.53
U	H	F	SI	POS	-7.47	-1.19	14.97

Second, the synthesized f_0 of the post-focus words in Figure 5e and 5f is near-identical to the natural f_0. The cases in which the synthesized f_0 is lower than the natural f_0 as in Figure 5b, 5e, 5h, 5i, 5k, 5l, 5n, 5o, 5q and Figure 5r are more than the cases in which the synthesized f_0 is higher than the natural f_0 as in Figure 5t and 5x.

Third, the synthesized f_0 of the pre-focus words are mostly similar to the natural f_0. This is clearly visible in most of the f_0 contours in which the focused word is sentence-penultimate or sentence-final as in Figure 5ai, 5ai. In the pre-focus region, the synthesized f_0 of the sentence-initial word shows no deviation from its natural counterpart. Indeed, the f_0 deviation in pre-focus region is less than that observed in the post-focus region or on-focus region discussed earlier.

Fourth, the synthesized f_0 of the neutral-focus utterances show a mixed picture. In Figure 5y, 5ab, 5ae, and Figure 5ah, the synthesized f_0 is near-identical to the natural f_0. In Figure 5a, 5d, 5g and Figure 5j, the synthesized f_0 appear different from the natural f_0, although the difference is less in Figure 5m, 5p, 5s and Figure 5v.

Finally, in the modelling, declination was not included as a function; yet it was nevertheless predicted, as can be seen in Figure 5b, 5c, 5d, 5e, 5f, 5g, 5h, 5m, 5n, 5o, 5p, 5r, 5s, 5t, 5ac, 5ad, 5ae, 5af, 5ai and Figure 5aj. This seems to provide support for the observation that there is no strong need to explicitly model declination in languages such as English (Xu and

Prom-on, 2014).

Overall, the graphical comparison in this section has shown that the multi-functional pitch targets extracted by PENTAtrainer from the nine speakers can accurately predict natural f_0 of the two speakers not included in the model training, including, in particular, the on-focus expansion and post-focus compression of f_0. The closeness of fit of the f_0 contours further suggests that the other putative categories included in the modelling, i.e., Stress, Weight, Position of word-level stress (Sposition), and Prosodic word (Pword) are also helpful for generating f_0 contours that resemble those of natural EA intonation.

3.2.2. Effect of Individual Annotated Functions

This section examines the contribution of each of the annotated functions listed in (16) by removing each of them at a time during both training and testing. Unlike in the previous section, however, here the synthetic f_0 is compared to the original f_0 of the nine speakers used in the model training. This is to avoid confounds due to properties peculiar to the two speakers not included in the model training. Table 9 shows the synthesis accuracies when different combinations of the functions were imposed. A smaller RMSE (root mean square error) indicates a better local f_0 fit between the synthetic and original f_0 (Xu and Prom-on, 2014). RMSE measures the difference at every time-normalised f_0 point between the synthetic and original f_0 contours, and thus provides an index of the average mismatch of the contours (Xu and Prom-on, 2014). A higher correlation indicates a better match of the overall f_0 contours.

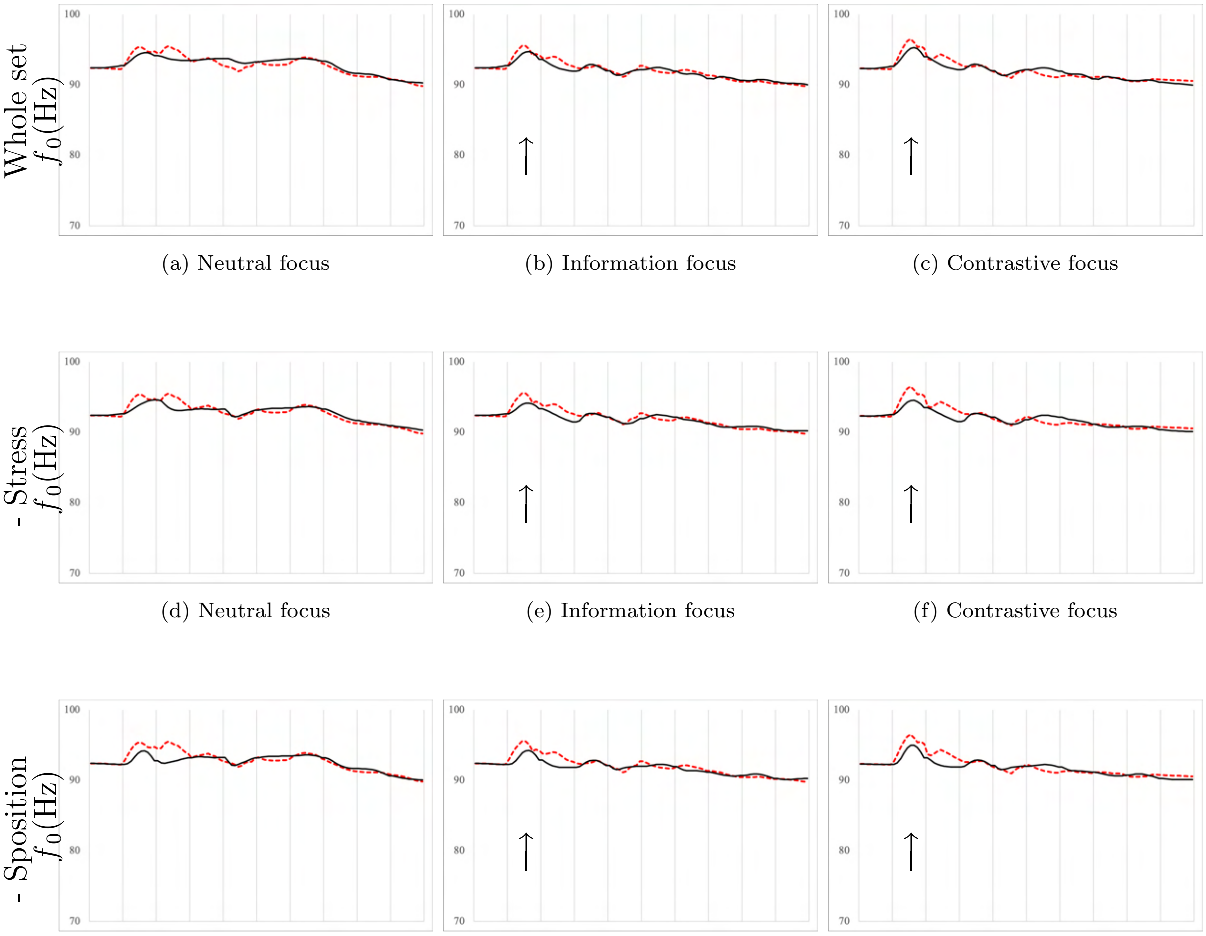
Table 9: Mean RMSEs (in st) and correlation coefficients of the synthesized f_0 against the original, with different functional combinations, averaged across 9 speakers. “Whole set” means all 5 functions were included in learning and synthesis, and “–Stress” means the stress function was excluded, and so on. The value of the standard deviation is between parenthesis.

Functions	**RMSE** (in st)	**Correlation**
Whole set	1.80 (0.95)	0.71 (0.21)
–Stress	1.78 (0.95)	0.70 (0.22)
–Sposition	1.78 (0.97)	0.69 (0.21)
–Focus	1.84 (0.94)	0.67 (0.22)
–Weight	1.83 (0.93)	0.66 (0.22)
–Pword	1.87 (0.97)	0.63 (0.23)

In Table 9, the Whole set condition shows the highest correlation of 0.71, while all the other conditions have lower correlations. This indicates that the removal of any single function reduced global f_0 fit relative to the Whole set condition. In terms of RMSE, in contrast, –Stress and –Sposition actually showed better local fit than the Whole set condition. One possible reason is that RMSE values less than 2 semitones are already very good compared to modelling results in previous studies on Mandarin (RMSE of 2.72 st and correlation of 0.87) and English (RMSE of 2.77 st and correlation of 0.77) (Xu and Prom-on, 2014; Liu et al., 2015). Therefore, the small fluctuations between 1.78 st and 1.87 st across all the conditions may not be very meaningful. On the other hand, there is an apparent drop in performance based on both RMSE and correlation in the –Focus, –Weight and –Pword conditions, indicating that their contributions to the good f_0 fit in the Whole set condition may be the most important.

To find out more details about the contribution of individual functions, mean time-normalised synthetic and original f_0 contours of the sentence /Ma.**lī**.kah **ḥa**.mat **La.**ma min **Mu.**na/ 'Malīkah protected Lama from Muna' in different focus conditions are displayed in Figure 6. These mean f_0 contours are averaged across all repetitions by all the nine speakers included in the model training. The synthesized f_0 contours in each row were generated either with all the five functions, or with one of the functions missing (See Appendix D for the parameters extracted from all the training conditions).

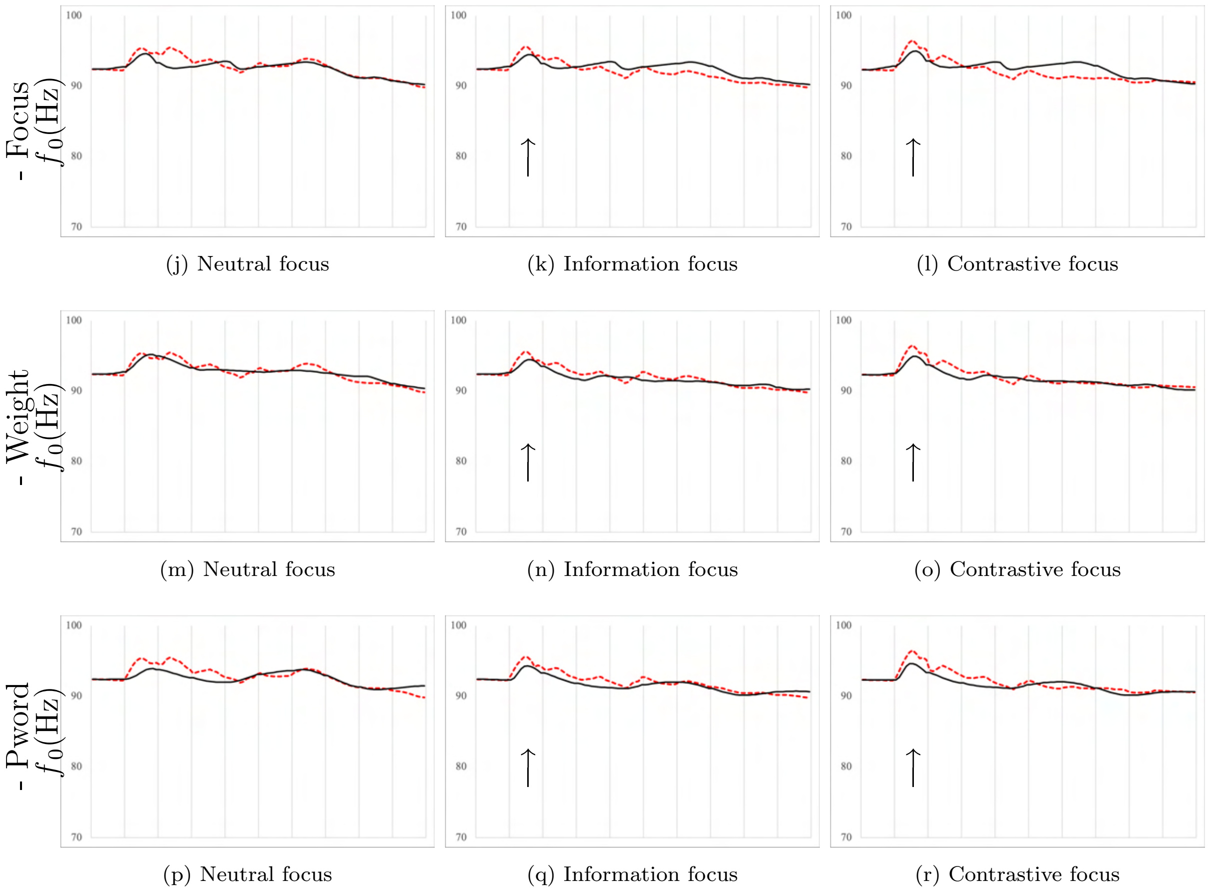
From the plots in Figure 6, three observations can be made. First, the f_0 fit in the Whole set condition, particularly the structures with information focus and contrastive focus, seems to agree better with the ranking based on correlation than the ranking based on RMSE in Table 9. Second, there are clear effects of missing focus in Figure 6k and 6l, where the synthetic f_0 no longer shows post-focus compression. Finally, in contrast to the –Focus condition, missing any of the other functions did not result in drastically different post-focus f_0, despite the changes in RMSE and correlation shown in Table 9.



(a) Neutral focus (b) Information focus (c) Contrastive focus

(d) Neutral focus (e) Information focus (f) Contrastive focus

(g) Neutral focus (h) Information focus (i) Contrastive focus



(j) Neutral focus

(k) Information focus

(l) Contrastive focus

(m) Neutral focus

(n) Information focus

(o) Contrastive focus

(p) Neutral focus

(q) Information focus

(r) Contrastive focus

Figure 6: Mean time-normalised original (red dotted line) and synthetic (black solid line) f_0 contours of an example "Ma.**lī**.kah **ḥa**.mat **La.**ma min **Mu.**na". The vertical lines mark the syllable boundaries. The upper arrow indicates the position of the stressed syllable of on-focus words. Each row corresponds to the results of implementing each functional layer: including all sets, excluding the stress layer (-Stress), excluding the word-level stressed syllable position layer (-Sposition), excluding the focus layer (-Focus), excluding the syllable weight layer (-Weight), and excluding the prosodic word layer (-Pword).

Combining f_0 contour comparisons in Figure 6 to the RMSE and correlation values in Table 9, some further observations can be made. First, excluding the Stress function resulted only in a very small drop in correlation and even a small improvement in RMSE. This suggests that the missing contribution of stress to the overall f_0 was largely made up for by the other annotated functions, particularly Weight and Pword. On the other hand, –Weight and –Pword both show sizeable increase of RMSE and decreases of correlation, indicating that each of them may have contributed something additional in the Whole set condition. More importantly, the deviations from the original f_0 in the –Sposition, –Weight and –Pword conditions all occur in the pre-focus portions of an utterance. This means that these functions are all about different degrees of stress/pitch accent, etc., while post-focus compression, the critical component of focus is fully captured by the Focus function. Finally, the low correlation compared to previous modelling studies (Xu and Prom-on, 2014; Liu et al., 2015) suggest that there may be greater tolerance for cross-speaker variability in terms of the exact pitch targets related to word stress. This is supported by the high naturalness rating by the listening subjects to be presented next.

Overall, the exploration of the individual functions in this study has produced some interesting preliminary findings. First, the disappearance of post-focus compression when and only when focus was excluded during model training seems to support the importance of focus as an independent prosodic function in EA. Second, for stress, which is jointly predictable in EA by syllable weight and syllable position in word according to (9), its exclusion from model training did not reduce f_0 fitting more than Weight, Pword or Sposition. This seems to offer modelling evidence that stress in EA is not independent of the other co-varying factors, but exactly which of these factors are the most essential needs to be tested in future studies. Third, the sizeable drops in f_0 fitting in the –Weight and –Pword conditions demonstrate that these two factors themselves may need to be taken seriously in the modelling of Arabic intonation. Finally, the unclarity of the role of these stress-related factors from the modelling results seems to have pointed to the very core of the problem, namely, the predictability of stress as stated in (9) makes it hard to experimentally control stress in EA independently of its predicting factors. In other words, the severity of a problem which is otherwise purely theoretical is highlighted by computational modelling.

3.3. Perception Experiment

In this section, we present the results of the perception experiment on focus recognition and naturalness judgement for the PENTAtrainer generated synthetic f_0 contours. The goal is to show whether functional based computational modelling can sufficiently capture perceptually relevant focus prosody, and whether the model can also capture sufficient details in other respects of f_0 that may affect the naturalness of synthetic prosody.

3.3.1. Materials

The stimuli used in the perception experiment were the same as those reported in 3.2.1. The f_0 contours were generated with the 26 sets of multi-functional targets learned from nine speakers (Table 8), and applied to the utterances of the two speakers not included in the training, with the original duration patterns of the two speakers, resulting in 360 synthetic utterances. In addition, the original utterances of the two speakers of the same sentences (without resynthesis) were used as the control. Each listening subject heard two versions of each of the 360 sentences: the original recording and the PENTAtrainer-synthesized sound.[14] Therefore, each subject heard a total of 720 sentences (2 speakers x 4 words x 3 foci x 3 sets of sentences x 5 repetitions x 2 types of sounds (i.e. original and synthetic) = 720) in random order.

3.3.2. Subjects

Twenty four female native speakers of Emirati Arabic served as subjects in this experiment. Before arriving at this final number, 107 participants were excluded: 42 participants failed to pass the headphone screening task (see §3.3.3), 11 participants did not learn to recognize the emphases with sufficient accuracy for our cut-off of 80% correct at the end of training (see §3.3.3). 54 participants did not complete the experiment. Participants were tested online using the Gorilla Experiment Builder (`https://gorilla.sc/`). They were all undergraduate students at United Arab Emirates University, born and raised in the city of Al Ain. They had no self-reported speech and hearing disorders and their ages ranged from 18 to 23 (*mean*

[14] PENTAtrainer is developed as a stand alone Praat script integrated with two Java programs, The resynthesis method is the Praat internal implementation of the PSOLA algorithm.

age = 21.3, *SD* = 2.1 years). None of them had served as a speaker in the production experiment, but they were comparable to the participants of the production experiment in age, gender and education.

3.3.3. Procedure

After providing informed consent, all participants completed the screening task to ensure that they were wearing headphones and could hear the sounds played to them (Woods et al., 2017). Following the headphone screening task, they were assigned to a training task. In the training task, they listened to 10 sentences produced by native speakers of EA from the production experiment. Their task during training was to identify (a) focus (initial, medial, final, or none, described to them as emphasis) and (b) whether the sound is a human sound or synthesized by a computer. The participants had to correctly answer 80% of the focus question before moving to the actual perception experiment. By the end of the training task, listeners were successful in focus identification. No feedback was given to the participants as to the accuracy of their focus/naturalness judgement.

In each trial, the participant listened to the target sentence, and judged (1) which word in each utterance, or none of them, was emphasized, and (2) whether the sound they heard was a human sound or computer-generated. The naturalness judgement was made immediately after the focus judgement. In each trial, the six response categories (initial focus, medial focus, final focus, no emphasis, human, made by computer) were displayed as choices on the computer screen, and the subject clicked on the one that matched her impression after hearing each sentence. The next sentence was played after a choice was made. The order of the sentences was fully randomised, but blocked by repetition. That is, one full repetition of all the sentences were finished before the start of another repetition, with a short break in between. The whole process took about 45 minutes on average. The two tasks as presented to the participants are shown in Appendix E.

3.3.4. Results

A. Focus Judgement

Table 10 is the confusion matrix showing the rates of recognition of the four focus categories (column) divided by the original and synthetic focus conditions (row). Overall, native listeners were able to identify focus from both the original and synthetic stimuli most of the time, although the original utterances had significantly higher accuracy than the synthetic ones. To investigate that further, a series of Linear Mixed-Effects models were applied to analyse the recognition accuracy in percentage, using the lme4 package (Bates et al., 2015) in R (R Core Team, 2019). The model included three potential fixed effects and their interactions: prosody source (original and synthetic), focus type (information focus, contrastive focus, neutral focus), and focus location (initial, penultimate, final). We included the fixed effects and their interactions only if it was judged to be superior to a less fully specified model. The random intercepts for participant/listener and by-participant/listener random slopes for the fixed effects were included maximally (Barr et al., 2013). The *p* values were obtained by likelihood ratio tests of models with and without the fixed effects and the interactions (R codes used are in Appendix F).

Table 10: Confusion matrix of focus perception (percentage). *None* refers to neutral focus. Red indicates higher values, while blue indicates lower values. The intensity of the colour represents the magnitude of the value. The darker the color is, the more extreme the value becomes.

		Target Word			
Focus Conditions	**Type of sound**	**None**	**Initial**	**Penultimate**	**Final**
Neutral Focus	Original	22%	36%	38%	4%
	Synthetic	37%	42%	19%	3%
Initial Information Focus	Original	15%	80%	5%	1%
	Synthetic	18%	75%	4%	4%
Initial Contrastive Focus	Original	4%	94%	2%	0%
	Synthetic	12%	83%	5%	1%
Penultimate Information Focus	Original	8%	14%	79%	0%
	Synthetic	20%	13%	65%	2%
Penultimate Contrastive Focus	Original	10%	12%	76%	2%
	Synthetic	20%	18%	61%	2%
Final Information Focus	Original	21%	18%	53%	8%
	Synthetic	40%	26%	30%	5%
Final Contrastive Focus	Original	23%	22%	43%	13%
	Synthetic	45%	31%	18%	8%

We obtained the following results. First, the original utterances were interpreted with significantly higher accuracy than the synthetic ones (χ^2 =10.206, df = 1, $p < 0.001$). Second, the accuracy of recognizing contrastive focus was significantly higher than that of recognizing information focus (χ^2 =4.550, df = 1, $p < 0.032$). Third, focus location also had a significant main effect (χ^2 =73.785, df = 2, $p < 0.001$), in that initial and penultimate focus had significantly higher identification accuracy than final focus (initial vs. penultimate $p < 0.026$, initial vs. final $p < 0.001$, and penultimate vs. final $p < 0.001$). Fourth, the interaction between focus type and focus location was significant (χ^2 =1 1.782, df = 2, $p < 0.003$). The difference between the identification accuracy of contrastive focus and information focus was more pronounced in initial focus ($p < 0.004$) than in penultimate focus ($p = 0.857$) and final focus ($p = 0.814$). The interaction between prosody source and focus type (χ^2= 0.673, df = 1, p = .412) was non-significant. Here, it is interesting to note that the synthetic stimuli were generated with no distinction between information focus and contrastive focus, and so the only source of difference between the two focus types would have come from the duration patterns, which are kept intact during synthesis. The interaction between prosody source and focus location (χ^2= 5.399, df = 2, p= 0.067) was also non-significant. Table 10 shows that the accuracy was poorer for synthetic utterances, especially in initial contrastive focus and both penultimate focus conditions, compared to the original utterances. Finally, the three-way interaction was non-significant (χ^2 = 6.789, df = 5, p = 0.237).

Overall, the results of the perception experiment show that listeners were able to identify focus from both the original and the synthetic stimuli most of the time, although the original utterances had significantly higher accuracy. This indicates that there is still room for further improvement in our computational modelling. However, it would also be interesting to see whether other computational approaches could reach a similar level of performance in future studies.

The data analyzed in this section and the reproducible analysis scripts are publicly available at Open Science Framework (OSF) (`https://osf.io/w5qvh`).

B. Naturalness Judgement

This section reports the results from the perception experiment in which the participants heard a set of synthesized and unsynthesised stimuli to test the human-likeness of the output of PRAAT's PSOLA function.[15] Figure 7 shows the percentage of judging synthetic sounds and original sounds as natural. Almost all original sounds were judged by the native listeners as natural sounds. As for the synthetic sounds, the highest natural sound judgement was associated with neutral focus (84%), followed closely by initial information focus (78%) and final information focus (77%), with penultimate information focus being the lowest (71%).

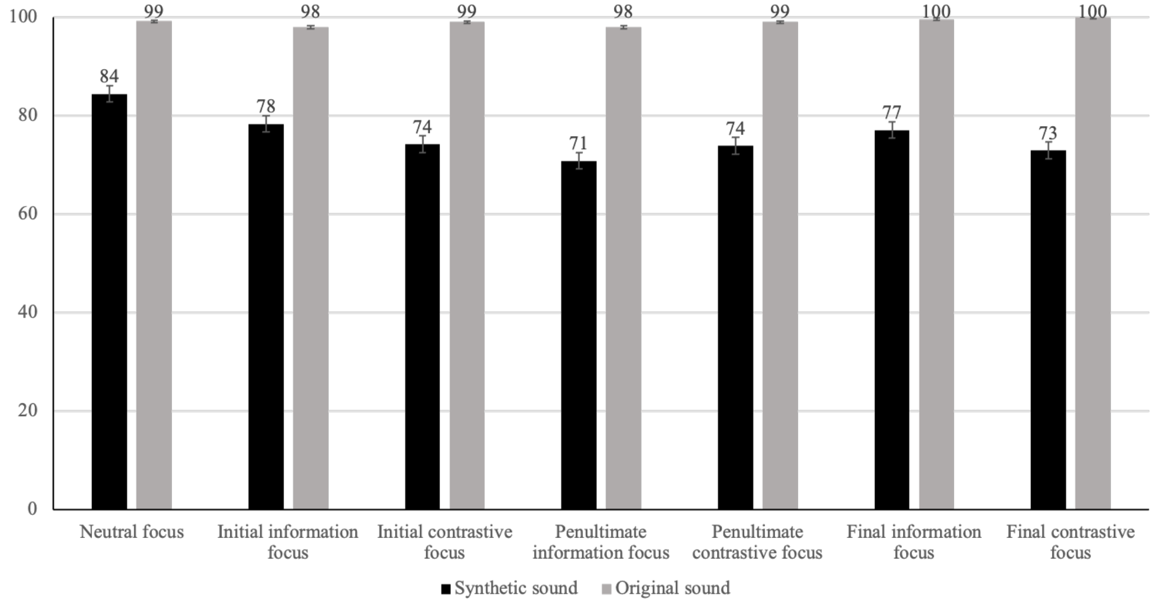


Figure 7: Naturalness perception (in percentage %). Error bars represent standard errors.

Splitting the data by individual listeners in Table 11, we can see great variability in how proficient individuals are at recognizing synthetic sounds. Eight listeners (p1, p2, p6, p8, p10, p17, p20 and p22) judged all synthetic sounds as natural utterances, and none of the listeners judged all the synthetic sounds as synthetic ones. Only participant p21 correctly identified synthetic sounds 93% of the time, which is the highest recognition rate among the listeners. Overall, approximately one-third of EA listeners judged all the synthetic sounds as natural.

Table 11: Percentage of "naturalness" responses to synthetic-sound stimuli from each listener. Info and Cont stand for information focus and contrastive focus respectively. *Info* and *Cont* refer to information focus and contrastive focus respectively.

	p1	**p2**	**p3**	**p4**	**p5**	**p6**	**p7**	**p8**	**p9**	**p10**
Neutral Focus	100	100	97	67	100	100	80	100	83	100
Initial Info.	100	100	100	30	90	100	100	100	90	100

Continued on next page

[15] Note that we report the results for the original stimuli for the purpose of using natural speech as a control, which differs from the popular practice of using a generic (often low quality) synthesis as the baseline. Using natural speech as baseline has been used in PENTAtrainer-based research since Prom-on et al. (2009).

Table 11 – continued from previous page

Initial Cont.	100	100	100	30	100	100	100	100	100	100
Penultimate Info.	100	100	100	50	90	100	70	100	20	100
Penultimate Cont.	100	100	100	20	90	100	70	100	80	100
Final Info.	100	100	100	70	90	100	80	100	60	100
Final Cont.	100	100	100	20	100	100	70	100	60	100
	p11	**p12**	**p13**	**p14**	**p15**	**p16**	**p17**	**p18**	**p19**	**p20**
Neutral Focus	100	83	67	100	57	10	100	93	97	100
Initial Info.	50	70	50	80	60	20	100	80	80	100
Initial Cont.	10	60	50	80	40	0	100	40	100	100
Penultimate Info.	100	60	50	100	0	0	100	40	100	100
Penultimate Cont.	100	90	60	100	10	0	100	20	100	100
Final Info.	100	60	70	100	30	10	100	80	100	100
Final Cont.	100	70	80	100	0	0	100	30	100	100
	p21	**p22**	**p23**	**p24**						
Neutral Focus	7	100	80	100						
Initial Info.	0	100	100	100						
Initial Cont.	0	100	80	100						
Penultimate Info.	0	100	20	100						
Penultimate Cont.	0	100	30	100						
Final Info.	0	100	10	90						
Final Cont.	0	100	30	90						

3.4. Discussion

The above results demonstrate that intonation of EA can be effectively modelled computationally with PENTAtrainer. Using only 26 parameters (Table 8) extracted from 1620 EA utterances in the learning dataset, f_0 contours of 360 utterances were generated. The number of parameters used for EA data was lower than the number of parameters used for other languages, e.g. 30 parameters for 2500 Thai disyllabic phrases, 84 parameters for 1289 Mandarin utterances and 78 parameters for 960 English utterances (Xu and Prom-on, 2014). It is also lower than the number of parameters used to predict double focus in American English (146) (Liu et al., 2015).

The computational modelling presented in this section provided the answers to the research questions in (10c and 10d), repeated below for convenience.

(17) a. Can Emirati Arabic focus intonation be computationally captured by an articulatory-functional model?
> Yes.

b. What can computational modelling reveal about Emirati Arabic intonation that we cannot easily learn from acoustic analysis alone?
> Not only focus, but also Weight, Pword, Sposition and Stress are important factors that jointly shape the intonation of Emirati Arabic. The high correlation of 0.8 and low RMSE 1.93 in the Whole set condition suggests the importance of including the other putative functions in the modelling of EA intonation.

4. General Discussion and Conclusion

We have analyzed focus intonation of Emirati Arabic and modelled it computationally with PENTAtrainer (Prom-on et al., 2012; Xu and Prom-on, 2014), based on the theoretical

PENTA model (Xu, 2005; Xu et al., 2015, 2022). We first performed systematic acoustic analyses of prosodic focus to find out how focus is intonationally encoded in EA. The results showed that focus in EA is encoded by an expansion of the pitch range in the on-focus region, a reduction of pitch range in the post-focus region, and a reduction of excursion size in the pre-focus region. These results established Emirati Arabic as a language with post-focus compression (+PFC), hence grouping it with the other Arabic dialects examined so far, including Egyptian Arabic (Hellmuth, 2006*b*, 2009), Hijazi (Alzaidi et al., 2019) and Lebanese Arabic (Chahal, 2001). However, Makkan Arabic as examined by Alzaidi (2022) is different from all the other Arabic dialects. This is because Makkan Arabic as reviewed in §1.1.2 is without post-focus compression. Interestingly, the present analysis showed a significant pre-focus reduction of duration and pitch excursion size. The excursion size reduction is similar to Lebanese Arabic (Chahal, 2001), but different from Egyptian (Hellmuth, 2006*b*, 2009), Hijazi Arabic (Alzaidi et al., 2019) and Makkan Arabic (Alzaidi, 2022).

Another finding of the acoustic analysis was that the difference between contrastive focus and information focus lies mainly in the on-focus words. In contrast, the acoustic measurements of contrastive focus and neutral focus were significant for not only on-focus words but also for post-focus and pre-focus words. This differs from the finding for Hijazi Arabic that f_0 excursion size and mean f_0 of the focused word are both greater in contrastive focus than in information focus as well as neutral focus (Alzaidi et al., 2019). Note that the method of eliciting contrastive focus in the present study is the same as in Alzaidi et al. (2019), the difference between the two focus types is smaller here. It was proposed in Alzaidi et al. (2019) that the significantly greater on-focus pitch range expansion found in the Hijazi contrastive focus was due to an incredulity effect elicited by the anecdotes presented to the participants in each trial. The incredulity implies an element of surprise, which involves increased f_0 beyond the level required by focus (Liu et al., 2021). In the present study, although similar anecdotes were used, their effects may have been limited by the formality of the recording situation. In Alzaidi et al. (2019), the recordings were made in a quiet room in the homes of the participants, which yielded a relaxing and familiar speaking environment. In the present study, the recordings were made in a laboratory, and most participants were not well acquainted with the experimenter. Both factors may have reduced the likelihood of eliciting sufficient incredulity from the participants. More research is needed to further investigate this issue.

The computational modelling of Arabic intonation in this study is the first of its kind for any Arabic dialect. The goal was to test, first, whether the tri-zone categorical representation of focus prosody can effectively generate continuous f_0 contours, as would be done in actual speech production, and second, what other linguistic categories also need to be incorporated in order to generate f_0 contours that closely resemble those of natural speech. The model applied was PENTA, which is composed of a target approximation model that simulates basic articulatory dynamics, and a parallel encoding scheme that simulates concurrent encoding of multiple functions (Xu, 2005; Xu et al., 2015, 2022). In the modelling experiment, we used PENTAtrainer (Prom-on et al., 2012; Xu and Prom-on, 2014) to learn syllable-sized pitch targets defined by focus and a number of other, putative functions from one set of utterances from the production study. The learned targets were applied in PENTAtrainer to generate f_0 contours that were then evaluated by comparison with the original intonation and by perceptual tests with native listeners. The functions included were Focus, Stress, Weight, Position of word-level stress (Sposition), and Prosodic word (Pword). We evaluated the closeness of fit both when all the five functions were included, and when one of them was excluded. As shown in Table 9, the accuracy in terms of correlation was the highest when all functions were included. The exclusion test showed that, interestingly, the omission of Stress resulted in the least reduction compared to the Whole set condition, indicating that stress contributed the least to the generation of fully detailed f_0 contours. We interpret this as evidence that stress in EA is not independent of the other co-varying factors, but exactly which of these factors are the most essential needs to be tested in future studies. Extending this reasoning further, we recognize that Weight, Position of word-level stress and Prosodic word all refer to various properties of a word that help to distinguish it from other words, in addition to its segmental composition in terms of consonants and vowels. In other words, their effects on various aspects of the detailed f_0 contours all provide additional phonetic cues that help to enhance the identifiability of the word itself. In this sense, they could all be considered as part of a lexical function for distinguishing words. This would make them, as a group, not that different from lexical stress in languages like English and German. The only difference is that in the latter case, the various cues, including f_0, intensity and duration, work together to make a rather robust categorical contrast. In EA, as a language without free word stress, the contrast is more distributed and less categorical - but these (i.e., f_0, intensity and duration) nevertheless matter for the naturalness of intonation in synthetic prosody.

The goal of the perceptual evaluation of the model-generated f_0 contours was to assess how well PENTA-based modelling can capture all the relevant intonational details of natural speech in EA. The results show that synthetic intonation has achieved rather high level of performance. However, the results from the perception experiment are still a bit below the level of natural speech. The results are nevertheless encouraging, and have set up a standard that may be useful for future modelling research with PENTA as well as other intonational models.

Appendix A: Test Materials

Due to limited space, we present below the test materials used for the first set of target sentences (11a). The other two sets (11b) and (11c) followed the same concepts represented and adopted in the following set.

4.1. Neutral Focus

The following scenarios were used to trigger neutral focus. The target sentence was embedded in the question-answer paradigm. it is an answer to the neutral focus question /wiš ʔas-sālfah?/ '*What happened*?

- **Scenario** 1

لين ولما ومنى أخوات. منى تبا تضرب لما ، لكن لين حمتها.

Glossing:

Līn wa Lama wa Muna ʔah̬wāt. Muna taba tiḍrib Lama lākin Līn ḥamatha.
Līn and Lama and Muna sisters Muna wanted hit Lama but Līn protected.her

'Līn, Lama and Muna are sisters. Muna wanted to hit Lama, but Līn protected her.'

Target sentence 1

.لين حمت لما من منى.

Līn ḥamat Lama min Muna.
Lı̄n protected Lama from Muna

‘Lı̄n protected Lama from Muna.’

- **Scenario** 2

لينا ولما ومنى أخوات. منى تبا تضرب لما لكن لينا حمتها.

Glossing:
Līna wa Lama wa Muna ʔaẖwāt. Muna taba tiḍrib Lama lākin Līna ḥamatha.
Līna and Lama and Muna sisters Muna wanted hit Lama but Līna protected.her
‘Leena, Lama and Muna are sisters. Muna wanted to hit Lama, but Leena protected her.’

Target sentence 2

لينا حمت لما من منى.

Līna ḥamat Lama min Muna.
Leena protected Lama from Muna

‘Leena protected Lama from Muna.’

- **Scenario** 3

منال ولما ومنى أخوات. منى تبا تضرب لما ، لكن منال حمتها.

Glossing:
Manāl wa Lama wa Muna ʔaẖwāt. Muna taba tiḍrib Lama lākin Manāl ḥamatha.
Manal and Lama and Muna sisters, Muna wanted hit Lama but Manāl protected.her

‘Manal, Lama and Muna are sisters. Muna wanted to hit Lama, but Manal protected her.’

Target sentence 3

منال حمت لما من منى.

Manāl ḥamat Lama min Muna.
Manal protected Lama from Muna

‘Manal protected Lama from Muna.’

- **Scenario** 4

مليكه ولما ومنى أخوات. منى تبا تضرب لما لكن مليكه حمتها.

Glossing:
Malīkah Lama wa Muna ʔaẖwāt. Muna taba tiḍrib Lama lākin Malīkah ḥamatha.
Malikah and Lama and Muna sisters, Muna wanted hit Lama but Malikah

protected.her
‘Malikah, Lama and Muna are sisters. Muna wanted to hit Lama, but Malikah protected her.’

Target sentence 4

مليكه حمت لما من منى.

Malīkah ḥamat Lama min Muna.
Malikah protected Lama from Muna
‘Malikah protected Lama from Muna.’

4.2. Information Focus

The following scenarios were used to trigger information focus, realized at sentence-initial position. The target sentence was embedded in the question-answer paradigm. it is an answer to the wh- question /minu ḥama Lama min Muna?/ ‘*Who protected Lama from Muna*?

- **Scenario** 1

منى تبا تضرب لما ، لكن لين حمتها

Glossing:
Muna tiba taba tiḍrib Lama, laken Līn ḥamatha.
Lin wanted to hit Lama but Lin protected.her
‘Muna wanted to hil Lama, but Līn protected her.’

Target sentence 1

.لين حمت لما من منى.

Līn ḥamat Lama min Muna.
Līn protected Lama from Muna
‘Līn protected Lama from Muna.’

- **Scenario** 2

منى تبا تضرب لما ، لكن لينا حمتها

Glossing:

Muna tiba taba tiḍrib Lama, laken Līna ḥamatha.
Lin wanted to hit Lama but Līna protected.her
‘Muna wanted to hit Lama, but Līna protected her.’

Target sentence 2

.لينا حمت لما من منى.

Līn ḥamat Lama min Muna.
Līn protected Lama from Muna
‘Līna protected Lama from Muna.’

- **Scenario** 3

منى تبا تضرب لما ، لكن منال حمتها

Glossing:
Muna tiba taba tiḍrib Lama, laken Manāl ḥamatha.
Lin wanted to hit Lama but Manal protected.her
‘Muna wanted to hit Lama, but Manal protected her.’

Target sentence 3

.منال حمت لما من منى.

Manāl ḥamat Lama min Muna.
Manal protected Lama from Muna
‘Manal protected Lama from Muna.’

- **Scenario** 4

منى تبا تضرب لما ، لكن مليكه حمتها

Glossing:
Muna tiba taba tiḍrib Lama, laken Malikah ḥamatha.
Lin wanted to hit Lama but Malikah protected.her
‘Muna wanted to hit Lama, but Malikah protected her.’

Target sentence 4

.مليكه حمت لما من منى.

Malikah ḥamat Lama min Muna.
Malikah protected Lama from Muna
‘Malika protected Lama from Muna.’

4.3. Contrastive Focus

The following scenarios were used to trigger contrastive focus, realized at sentence-initial position. The target sentence was embedded in the question-answer paradigm. it is an answer to the wh- question /minu ḥama Lama min Muna?, Layān?/ ‘*Who protected Lama from Muna*? *Layān*?

- **Scenario** 1

لين ولما ومنى وليان أخوات، أكبرهم ليان وأصغرهم لين. منى تبا تضرب لما، لكن لين حمتها.

Glossing:
Līn wa Lama wa Muna wa Layān ʔaḫwāt. ʔakbarhum Layān wa ʔaṣġarhum
Līn and Lama and Muna sisters eldest Layān and youngest Līn Muna
Līn Muna taba tiḍrib Lama lākin Līn ḥamatha.
wanted hit Lama but Līn protected.her
‘Līn, Lama, Muna and Layān are sisters. The eldest is Layān and the youngest is Līn. Muna wanted to hit Lama, but Līn protected her.’
Target sentence 1

لين حمت لما من منى.

Līn ḥamat Lama min Muna.
Līn protected Lama from Muna
‘Līn protected Lama from Muna.’

- **Scenario** 2

لينا ولما ومنى وليان أخوات، أكبرهم ليان وأصغرهم لينا. منى تبا تضرب لما، لكن لينا حمتها.

Glossing:
Līna wa Lama wa Muna wa Layān ʔaḫwāt. ʔakbarhum Layān wa ʔaṣġarhum
Līna and Lama and Muna sisters eldest Layān and youngest Līna Muna
Līna Muna taba tiḍrib Lama lākin Līna ḥamatha.
wanted hit Lama but Līna protected.her
‘Līna, Lama, Muna and Layān are sisters. The eldest is Layān and the youngest is Līna. Muna wanted to hit Lama, but Līna protected her.’

Target sentence 2

لينا حمت لما من منى.

Līna ḥamat Lama min Muna.
Līna protected Lama from Muna
‘Līna protected Lama from Muna.’

- **Scenario** 3

منال ولما ومنى وليان أخوات، أكبرهم ليان وأصغرهم منال. منى تبا تضرب لما، لكن منال حمتها.

Glossing:

Manāl wa Lama wa Muna wa Layān ʔaḫwāt. ʔakbarhum Layān wa ʔaṣġarhum
Manāl and Lama and Muna sisters eldest Layān and youngest Manāl Muna
Manāl Muna taba tiḍrib Lama lākin Manāl ḥamatha.
wanted hit Lama but Manāl protected.her
‘Manāl, Lama, Muna and Layān are sisters. The eldest is Layān and the youngest is Manāl. Muna wanted to hit Lama, but Manāl protected her.’

Target sentence 3

منال حمت لما من منى.

Manāl ḥamat Lama min Muna.
Manāl protected Lama from Muna
‘Manāl protected Lama from Muna.’

- **Scenario** 4

مليكه ولما ومنى وليان أخوات، أكبرهم ليان وأصغرهم مليكه. منى تبا تضرب لما، لكن مليكه حمتها.

Glossing:
Malikah wa Lama wa Muna wa Layān ʔaḫwāt. ʔakbarhum Layān wa
Malikah and Lama and Muna sisters eldest Layān and youngest Malikah
ʔaṣġarhum Malikah Muna taba tiḍrib Lama lākin Malikah ḥamatha.
Muna wanted hit Lama but Malikah protected.her
‘Malikah, Lama, Muna and Layān are sisters. The eldest is Layān and the youngest is Malikah. Muna wanted to hit Lama, but Malikah protected her.’

Target sentence 4

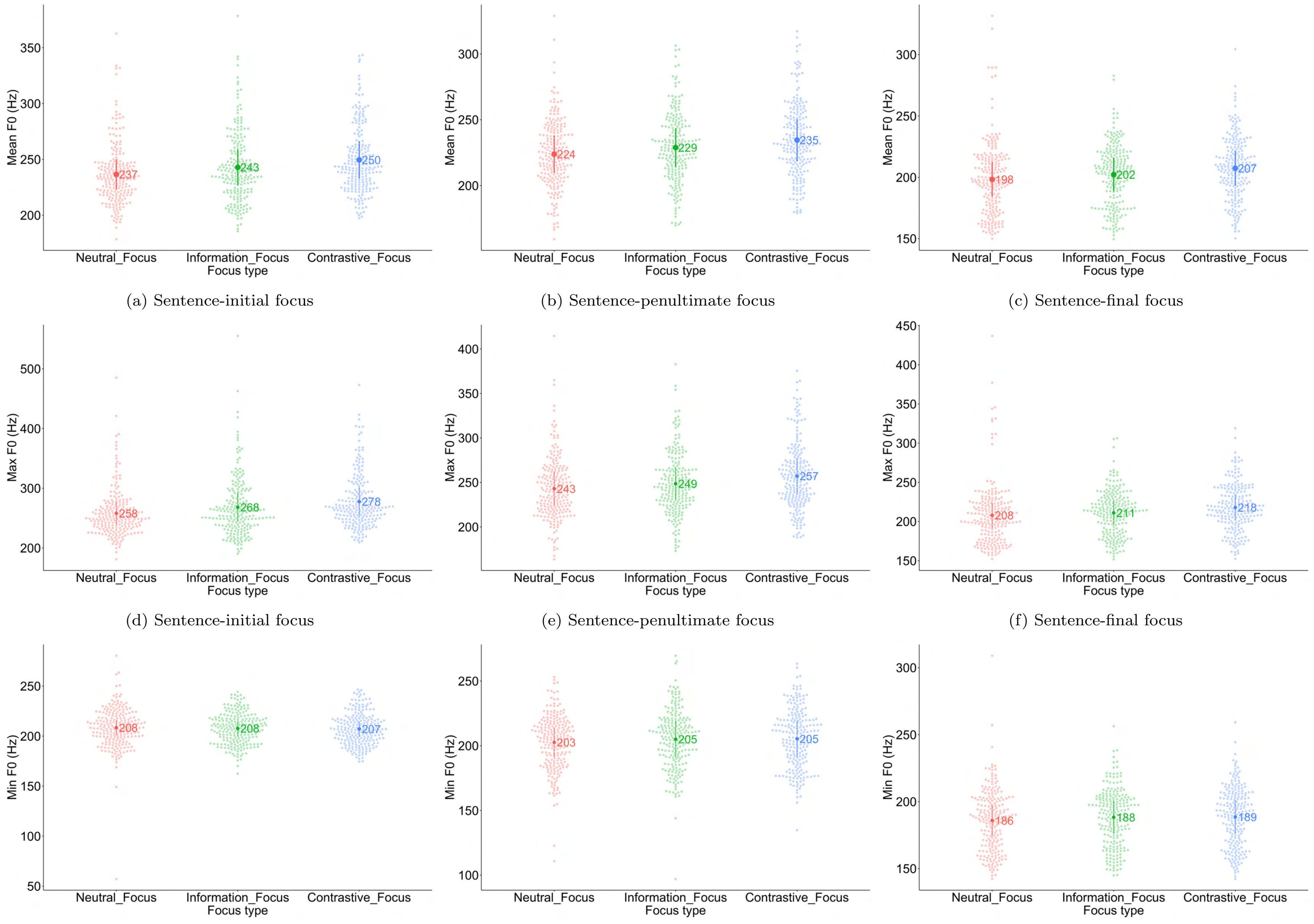
مليكه حمت لما من منى.

Malikah ḥamat Lama min Muna.
Malikah protected Lama from Muna
‘Malikah protected Lama from Muna.’

Appendix B: Samples of R Codes

```
model1<−lmer(maxf0~(Focus|Subject)+(Focus|Sentence)+
(Focus|Sentence:Subject),dat=mydat\) # baseline model

model2<− update(model1,.~.+Focus) # add main effect

#In the case of non−convergence
model2<− update(model1,.~.+Focus,
control=lmerControl(optimizer=”nloptwrap”,
optCtrl=list(algorithm=”NLOPT_LN_NELDERMEAD”)))


anova(model1,model2) # model comparison for obtaining p value


summary(glht(model2, linfct=mcp(Focus=”Tukey”)))
# example  of post−hoc comparison

emmeans(model2,pairwise ~Focus)
```

Appendix C: Model prediction plots of mixed effects model



(a) Sentence-initial focus

(b) Sentence-penultimate focus

(c) Sentence-final focus

(d) Sentence-initial focus

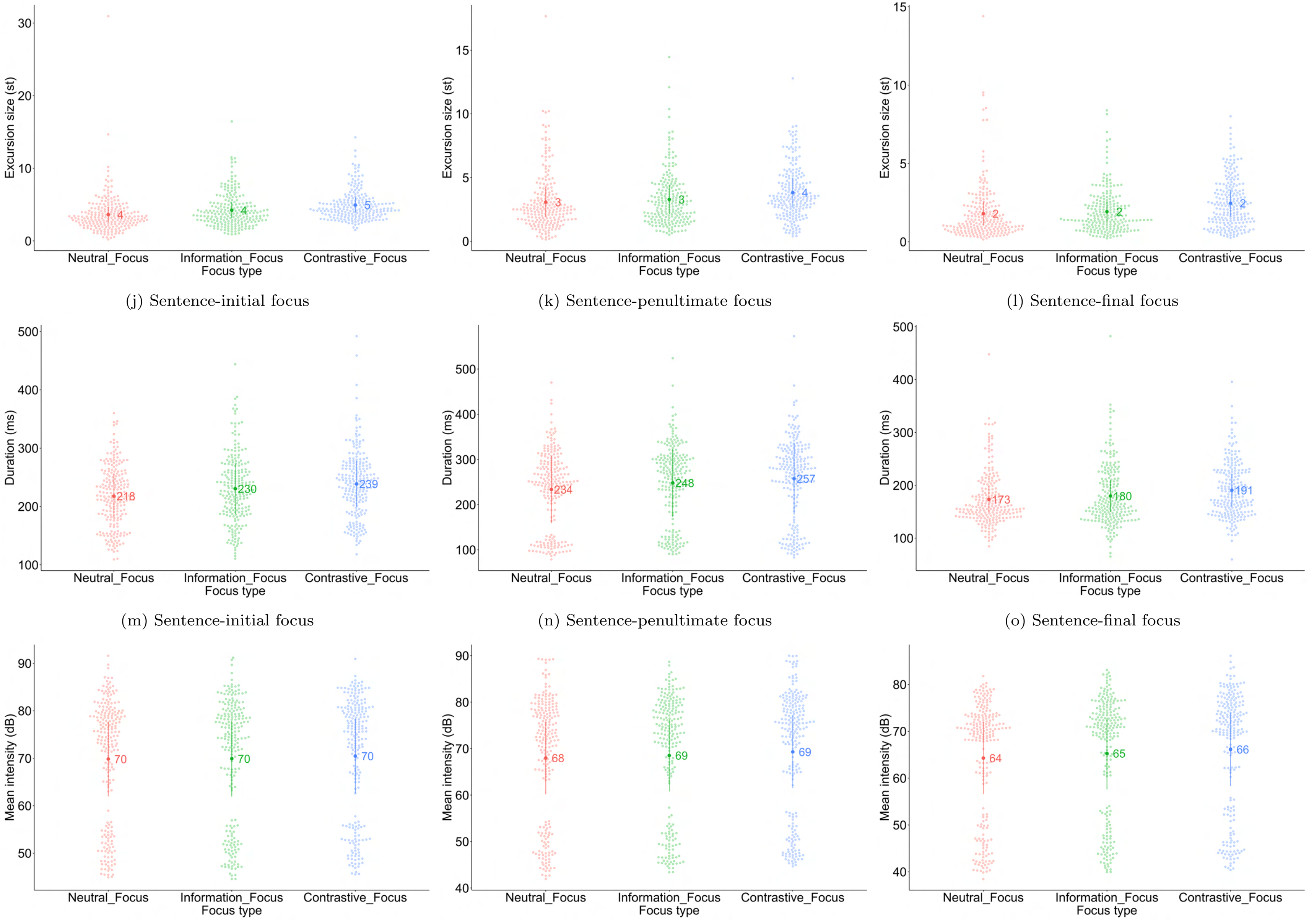
(e) Sentence-penultimate focus

(f) Sentence-final focus

(g) Sentence-initial focus

(h) Sentence-penultimate focus

(i) Sentence-final focus



(j) Sentence-initial focus

(k) Sentence-penultimate focus

(l) Sentence-final focus

(m) Sentence-initial focus

(n) Sentence-penultimate focus

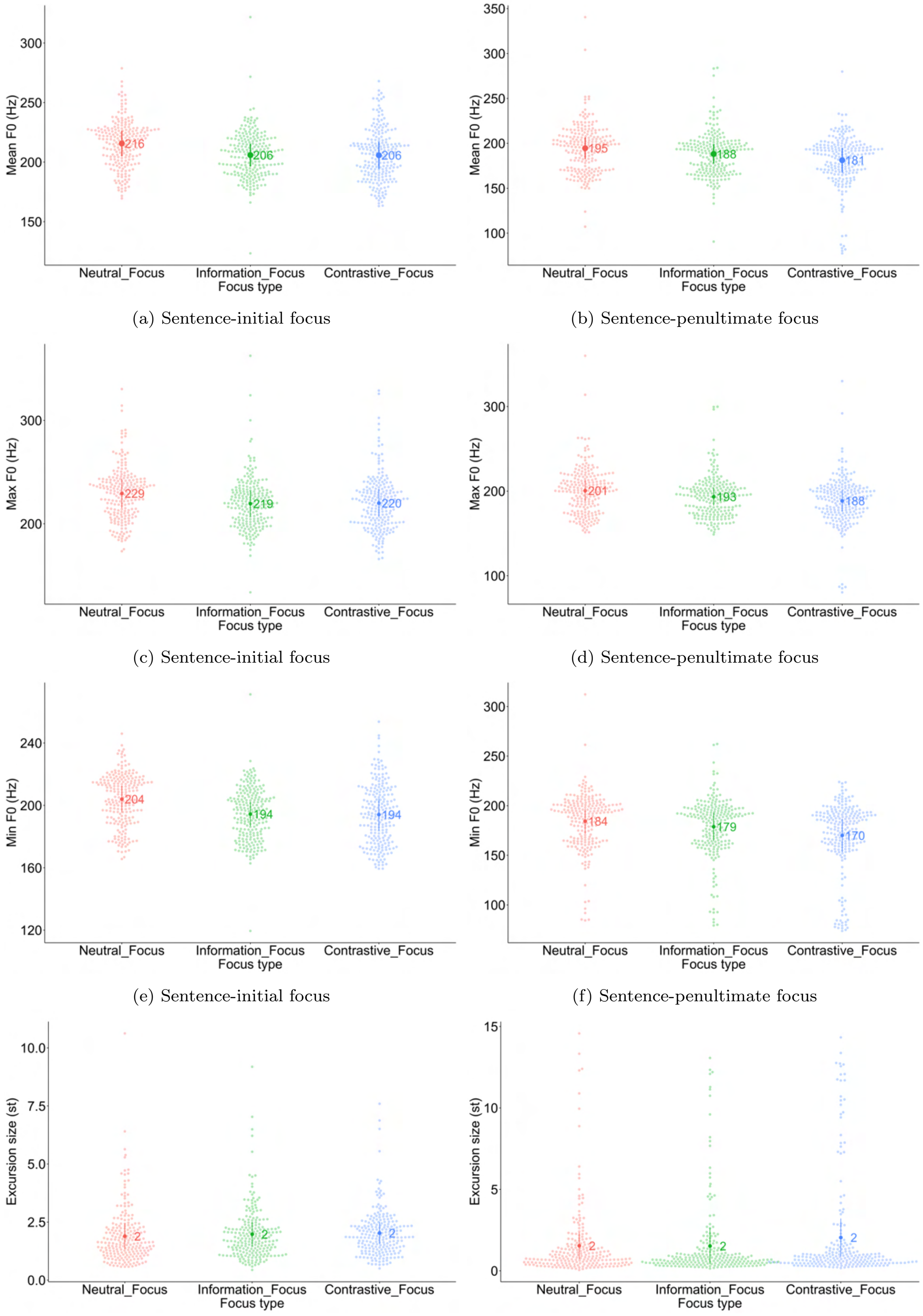
(o) Sentence-final focus

(p) Sentence-initial focus

(q) Sentence-penultimate focus

(r) Sentence-final focus

Figure 8: Model prediction plots of mixed effects model: On-focus region. The point ranges show the estimated marginal means (predicted values) and the lower/ upper bound of the 95% confidence interval for the predicted values. The beeswarm plot shows the individual data points by each participant in each sentence type.



(a) Sentence-initial focus

(b) Sentence-penultimate focus

(c) Sentence-initial focus

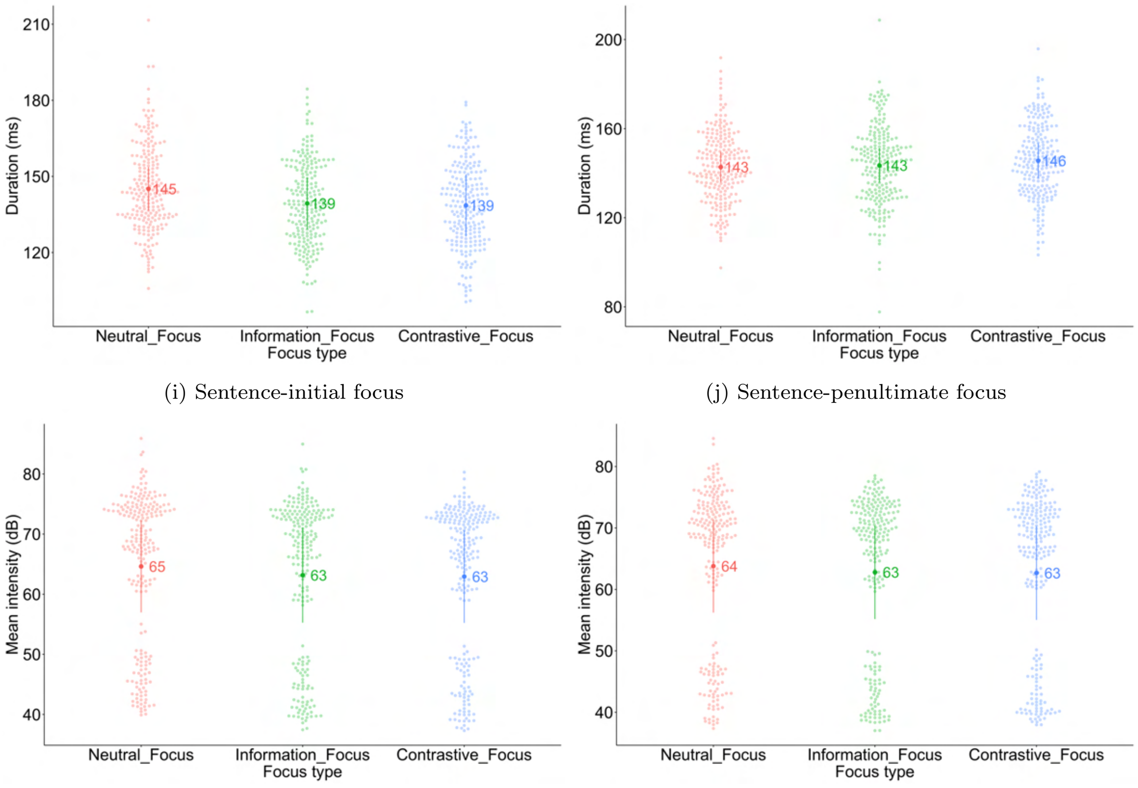
(d) Sentence-penultimate focus

(e) Sentence-initial focus

(f) Sentence-penultimate focus

(g) Sentence-initial focus

(h) Sentence-penultimate focus



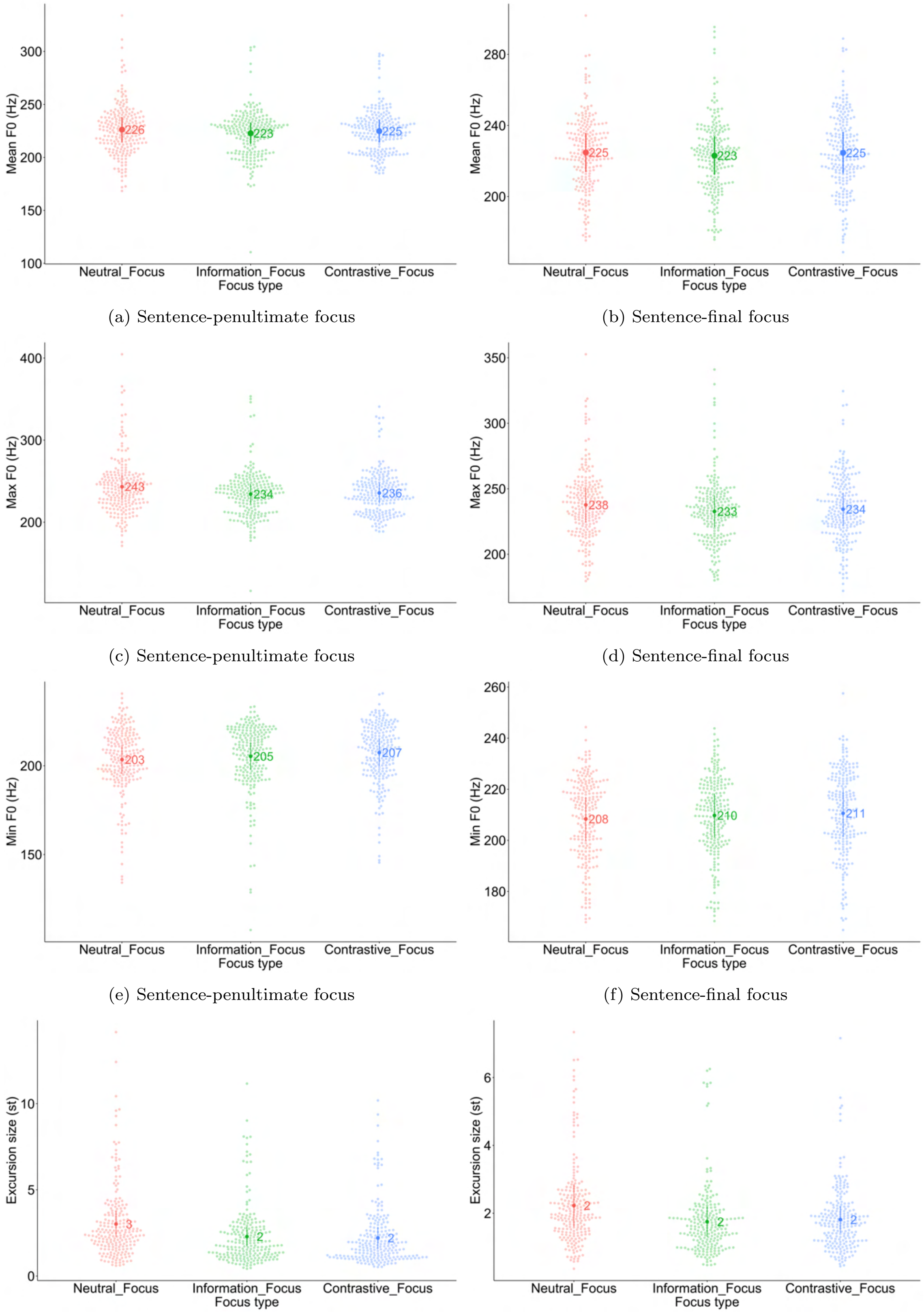
(i) Sentence-initial focus

(j) Sentence-penultimate focus

(k) Sentence-initial focus

(l) Sentence-penultimate focus

Figure 9: Model prediction plots of mixed effects model: Post-focus region. The point ranges show the estimated marginal means (predicted values) and the lower/ upper bound of the 95% confidence interval for the predicted values. The beeswarm plot shows the individual data points by each participant in each sentence type.



(a) Sentence-penultimate focus

(b) Sentence-final focus

(c) Sentence-penultimate focus

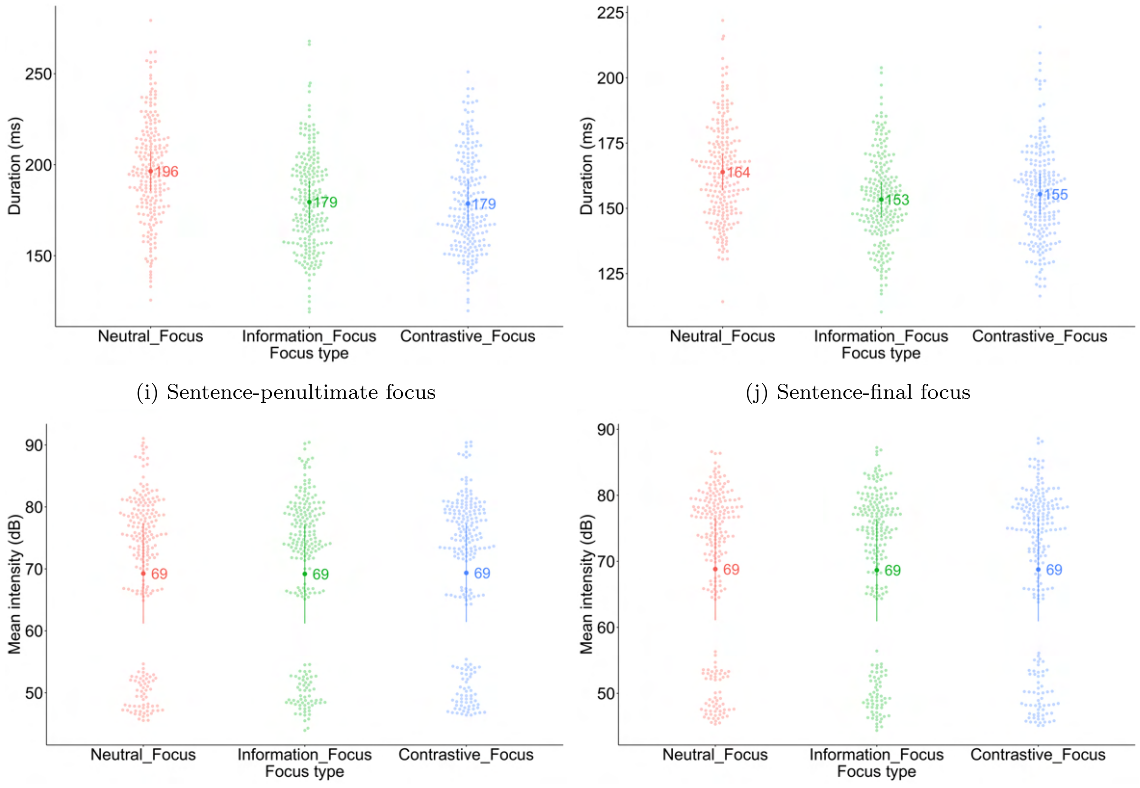
(d) Sentence-final focus

(e) Sentence-penultimate focus

(f) Sentence-final focus

(g) Sentence-penultimate focus

(h) Sentence-final focus



(i) Sentence-penultimate focus

(j) Sentence-final focus

(k) Sentence-penultimate focus

(l) Sentence-final focus

Figure 10: Model prediction plots of mixed effects model: Pre-focus region. The point ranges show the estimated marginal means (predicted values) and the lower/ upper bound of the 95% confidence interval for the predicted values. The beeswarm plot shows the individual data points by each participant in each sentence type.

Appendix D: Parameters

The following sets of parameters are extracted from the learning set that contains 1620 EA utterances from nine female native speakers of Emirati Arabic (presented in §3). F or stress function, *S* denotes stressed syllables and *U* denotes unstressed syllables. For weight, *R*, *H*, *L* denote superheavy, heavy and light syllables, respectively. For Sposition, *I*, *P* and *F* denote initial, penultimate and final position, respectively. For focus, *PRE*, *ON*, *POST* denote pre-focus, on-focus and post-focus regions, respectively.

Table 12: Parameters of EA intonation: -Focus condition. These parameters were used to produce the synthetic f_0 contours in Figure 6j, 6k and 6l

Stress	**Weight**	**Sposition**	**Pword**	**m** (st/s)	**b** (st)	λ
S	H	I	SI	-5.35	3.86	17.36
U	L	F	SF	-27.23	1.29	14.34
S	L	I	SI	46.86	-17.06	6.3
U	H	F	SF	5.46	-0.17	51.86
U	H	F	SI	67.2	28.91	2.83
S	L	I	SM	-9.66	-2.15	54.34
U	L	I	SI	13.19	-0.16	71.84
S	R	F	SF	-74.84	18.63	6.07
S	H	P	SM	-73.17	4.45	16.38

Continued on next page

Table 12 – continued from previous page

Stress	**Weight**	**Sposition**	**Pword**	**m** (st/s)	**b** (st)	λ
S	R	F	SI	-21.28	2.18	15.98
U	L	F	SI	-3	-1.25	100
S	H	I	SM	-31.7	-0.52	33.56
U	L	I	SM	-56.54	24.25	5.62

Table 13: Parameters of EA intonation: -Stress condition. These parameters were used to produce the synthetic f_0 contours in Figure 6d, 6e and 6f

Weight	**Sposition**	**Pword**	**Focus**	**m** (st/s)	**b** (st)	λ
H	I	SI	PRE	49.65	-2.8	15.16
L	F	SF	PRE	-88.33	15.58	9.86
L	I	SI	PRE	17.46	-0.77	23.87
H	F	SF	PRE	8.33	0.25	82.29
H	F	SI	PRE	91.25	-11.86	3.64
L	I	SM	PRE	-12.74	-1.48	19.08
R	F	SF	PRE	-8.31	2	10.57
H	P	SM	PRE	-23.3	6.27	9.42
R	F	SI	ON	-72.54	5.4	12.92
L	I	SI	POS	-12.22	-0.74	90.21
H	F	SF	POS	19.22	11.39	5.14
L	F	SF	POS	-20.53	-1.1	25.68
H	F	SI	POS	-22.9	1.6	15.26
L	I	SM	POS	-14.3	-2.27	25.79
H	I	SI	ON	-33.67	6.23	18.53
R	F	SF	ON	-29.99	0.24	17.48
H	P	SM	ON	-33.87	1.49	21.44
R	F	SI	PRE	-59.69	6.17	15.19
L	F	SI	PRE	0.39	-0.87	76.13
H	I	SM	PRE	-17.1	-1.28	41.89
L	I	SM	ON	-100	8.78	16.63
H	I	SM	ON	-53.77	0.82	23.93
L	I	SI	ON	55.73	30	1.93
L	F	SI	POS	-57.35	16.12	9.83

Table 14: Parameters of EA intonation: -Sposition condition. These parameters were used to produce the synthetic f_0 contours in Figure 6g, 6h and 6i

Stress	**Weight**	**Pword**	**Focus**	**m** (st/s)	**b** (st)	λ
S	H	SI	PRE	-19.25	5.61	16.91
U	L	SF	PRE	-33.37	2.73	16
S	L	SI	PRE	15.39	-0.54	25.54
U	H	SF	PRE	8.63	0.3	100
U	H	SI	PRE	3.89	-1.4	43.9
S	L	SM	PRE	41.59	-20.46	8.71
U	L	SI	PRE	7.65	-1.19	34.67
S	R	SF	PRE	-23.42	1.78	17.14
S	H	SM	PRE	-64.46	2.17	20.77
S	R	SI	ON	-46.84	1.59	21.23
S	L	SI	POS	-30.99	-0.53	46.35
U	H	SF	POS	-0.25	-0.86	34.99
U	L	SF	POS	-15.5	-1.21	28.34
U	H	SI	POS	63.93	30	3.34
S	L	SM	POS	-11.26	-2.61	51.51

Continued on next page

Table 14 – continued from previous page

Stress	**Weight**	**Pword**	**Focus**	**m** (st/s)	**b** (st)	λ
S	H	SI	ON	-16.33	5.06	18.68
S	R	SF	ON	-21.86	-0.17	25.73
S	H	SM	ON	-55.35	1.42	23.41
S	R	SI	PRE	-22.17	2.54	15.83
U	L	SM	PRE	-27.04	-6.22	7.01
S	L	SM	ON	-53.48	0.85	30.35
S	L	SI	ON	-32.96	6.51	17.76
U	L	SI	POS	-9.02	-2.15	100

Table 15: Parameters of EA intonation: -Weight condition. These parameters were used to produce the synthetic f_0 contours in Figure 6m, 6n and 6o

Stress	**Sposition**	**Pword**	**Focus**	**m** (st/s)	**b** (st)	λ
S	I	SI	PRE	-2.66	0.13	49.92
U	F	SF	PRE	-31.16	9.12	6.97
U	F	SI	PRE	67.79	14.41	1.64
S	I	SM	PRE	-9.91	-1.87	100
U	I	SI	PRE	96.88	-12.75	13.68
S	F	SF	PRE	-35.04	3.38	10.35
S	P	SM	PRE	-83.63	14.89	9.42
S	F	SI	ON	-21.95	0.5	45.34
S	I	SI	POS	-50.81	2.73	23.18
U	F	SF	POS	-10.24	-1.25	26.53
U	F	SI	POS	88.33	27.47	2.24
S	I	SM	POS	-62.65	3	15.87
S	I	SI	ON	-30.82	5.31	20.75
S	F	SF	ON	-16.67	-0.06	16.53
S	P	SM	ON	-44.67	2.19	19.92
S	F	SI	PRE	-29.35	3.26	15.23
U	I	SM	PRE	-100	-4.29	1.87
S	I	SM	ON	-53.15	16.53	5.88

Table 16: Parameters of EA intonation: -Pword condition. These parameters were used to produce the synthetic f_0 contours in Figure 6p, 6q and 6q

Stress	**Weight**	**Sposition**	**Focus**	**m** (st/s)	**b** (st)	λ
S	H	I	PRE	-17.5	3.21	13.01
U	L	F	PRE	-23.6	0.68	20.68
S	L	I	PRE	21.68	-4.36	10.47
U	H	F	PRE	7.96	0.48	6.49
U	L	I	PRE	21.79	-3.49	16.59
S	R	F	PRE	-48.85	5.4	14.28
S	H	P	PRE	-100	18.84	9.8
S	R	F	ON	-30.05	2.01	13.15
S	L	I	POS	-3.11	-1.48	15.95
U	H	F	POS	69.49	-23.5	5.53
U	L	F	POS	-12.97	-1.62	21.65
S	H	I	ON	36.71	-3.13	12.77
S	H	P	ON	-17.63	0.48	41.16
S	L	I	ON	-9.11	9.8	7.17

Appendix E: Perception experiment

*Screen task

Instructions in Arabic

ستسمع ثلاث أصوات مختلفة قم باختيار الصوت المنخفض جدافيهم عن طريق إختيار الخيار المناسب
من الثلاثة خيارات الموجودة.

Translation

You will hear three different sounds, choose the lowest-volume sound from these sounds.

*Training task and the main experiment test

Instructions in Arabic

ستسمع جملة ، قم بالإجابة على سؤالين السؤال الأول ماهي الكلمة المشدد عليها
صوتيا من بين الثلاث كلمات في الجملة ، وفي حالة كون كل الكلمات على مستوى واحد
من الصوت، قمت بإختيار خيار (غيز منبورة) اما السؤال الثاني فهل الصوت المسموع
صوتا حاسوبيا او صوت بشري

Translation

You will hear a sentence, choose which word in a sentence is emphasized. In case that all the sentence is produced at the same sound level, you choose "None". The second question is to choose whether the sound you hear is natural or made by computer.

Appendix F: R codes for perception experiment

```
#FocusRec represents recognition accuracy in percentage

model1<-lmer(FocusRec ~(ProsodyType|Participants)+
            (FocusType|Participants)+(FocusLocation|Participants),
            dat=dat_withFocus) # baseline model
model2<- update(model1,.~.+ProsodyType) # add main effects
model3<- update(model2,.~.+FocusType)
model4<- update(model3,.~.+FocusLocation)
model5<- update(model4,.~.+ProsodyType:FocusType) #add main effects
model5<- update(model4,.~.+ProsodyType:FocusLocation)
model5<- update(model4,.~.+FocusType:FocusLocation)
model6<- update(model5,.~.+ProsodyType:FocusType:FocusLocation)
anova(model1,model2) # to obtain p value by model comparison
summary(model5) # final model
emmeans(model5, pairwise~FocusType:FocusLocation) # post-hoc
comparison for the interaction effect
```

References

Al-Ani, S. (1992), 'Stress variation of the construct phrase in Arabic: A spectrographic analysis', *Anthropological Linguistics* **34**, 256–276.

Al Kaabi, M. and Ntelitheos, D. (2019), 'Rethinking templates: A syntactic analysis of verbal morphology in Emirati Arabic', *Glossa: a journal of general linguistics* **4**(1).

Alzaidi, M. (2014), Information structure and intonation in Hijazi Arabic, PhD Thesis, University of Essex.

Alzaidi, M. (2021*a*), 'F0 Peak Alignment, F0 Peak Location, and Focus Perception in Taif Arabic', *International Journal of Linguistics* **13**(6), 140–162.

Alzaidi, M. (2021*b*), 'Pitch Accent Distribution and Focus Structure in Taifi Arabic: A Production Study', *International Journal of English Linguistics* **12**(1), 179–196.

Alzaidi, M. S. (2022), 'Makkan arabic does not have post-focus compression: a production and perception study', *Phonetica* **79**(3), 247–308.

Alzaidi, M. S., Xu, Y. and Xu, A. (2019), 'Prosodic encoding of focus in Hijazi Arabic', *Speech Communication* **106**, 127–149.

Anderson, M., Pierrehumbert, J. and Liberman, M. (1984), Synthesis by rule of English intonation patterns, *in* 'ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 9, IEEE, pp. 77–80.

Aoun, J. E., Benmamoun, E. and Choueiri, L. (2009), *The syntax of Arabic*, Cambridge University Press.

Bani Younes, M. (2020), The role of phonology in the disambiguation of disjunctive questions, PhD thesis, University of York.

Barr, D. J., Levy, R., Scheepers, C. and Tily, H. J. (2013), 'Random effects structure for confirmatory hypothesis testing: Keep it maximal', *Journal of Memory and Language* **68**(3), 255–278.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015), 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* **67**(1), 1–48.

Beckman, M. E. and Pierrehumbert, J. B. (1986), 'Intonational structure in Japanese and English', *Phonology yearbook* **3**, 255–309.

Benkirane, T. (1998), 'Intonation in Western Arabic (Morocco)', *Intonation systems* pp. 345–359.

Blodgett, A., Owens, J. and Rockwood, T. (2007), 'An initial account of the intonation of Emirati Arabic', *Proceedings of the 16th ICPhS, Saarbruecken, Germany* pp. 1137–1140.

Boersma, P. and Weenink, D. (1992), 'PRAAT: Doing phonetics by computer', *www.praat.org* **Version 5.2.15**.

Botinis, A., Fourakis, M. and Gawronska, B. (1999), 'Focus identification in English, Greek and Swedish', *Proceedings of the 14th International Congress of Phonetic Sciences. San Francisco* pp. 1557–1560.

Breen, M., Fedorenko, E., Wagner, M. and Gibson, E. (2010), 'Acoustic correlates of information structure', *Language and Cognitive Processes* **25 (issue 7 and 9)**(7-9), 1044–1098.

Brown, G. and Hellmuth, S. (2022), 'Computational modelling of segmental and prosodic levels of analysis for capturing variation across Arabic dialects', *Speech Communication* .

Bruce, G. (1982), 'Developing the Swedish intonation model', *Working papers/Lund University, Department of Linguistics and Phonetics* **22**.

Brustad, K. (2000), *The syntax of Spoken Arabic*, Washington: Georgetown University Press.

Burdin, R. S., Phillips-Bourass, S., Turnbull, R., Yasavul, M., Clopper, C. G. and Tonhauser, J. (2015), 'Variation in the prosody of focus in head- and head/edge-prominence languages', *Lingua. International review of general linguistics. Revue internationale de linguistique générale* **165**, 254–276.

Cangemi, F., El Zarka, D., Wehrle, S., Baumann, S. and Grice, M. (2016), Speaker-specific intonational marking of narrow focus in Egyptian Arabic, pp. 335–339.

Chahal, D. (1999), A preliminary analysis of Lebanese Arabic intonation, *in* 'Conference of the Australian Linguistic Society', pp. 1–17.

Chahal, D. (2001), Modeling the intonation of Lebanese Arabic using autosegmental-metrical framework: a comparison with English, PhD thesis, University of Melbourne.

Chahal, D. (2003), 'Phonetic cues to prominence in Lebanese Arabic', *Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona* pp. 2067–2070.

Chahal, D. and Hellmuth, S. (2014), The intonation of Lebanese and Egyptian Arabic, *in* S.-A. Jun, ed., 'Prosodic Typology II: The Phonology of Intonation and Phrasing', Oxford University Press, pp. 365–404.

Chen, S.-w., Wang, B. and Xu, Y. (2009), Closely related languages, different ways of realizing focus, *in* 'Tenth Annual Conference of the International Speech Communication Association'.

Chen, Y. and Gussenhoven, C. (2008), 'Emphasis and tonal implementation in Standard Chinese', *Journal of Phonetics* **36**(4), 724–746.

Cooper, W. E., Eady, S. J. and Mueller, P. R. (1985), 'Acoustical Aspects of Contrastive Stress in Question–Answer Contexts', *The Journal of the Acoustical Society of America* **77**, 2142–2156.

Eady, S. and Cooper, W. (1986), 'Speech intonation and focus location in matched statements and questions', *Journal of the Acoustical Society of America* **80**(2), 402–415.

Eady, S. J., Cooper, W. E., Klouda, G. V., Mueller, P. R. and Lotts, D. W. (1986), 'Acoustical Characteristics of Sentential Focus: Narrow vs. Broad and Single vs. Dual Focus Environments', *Language and Speech* **29**(3), 233–251.

El Zarka, D. and Hödl, P. (2021), 'A study on the perception of prosodic cues to focus by Egyptian listeners: Some make use of them, but most of them don't', *Speech Communication* **132**, 55–69.

El Zarka, D., Kelterer, A. and Schuppler, B. (2020), An Analysis of Prosodic Prominence Cues to Information Structure in Egyptian Arabic., *in* 'Interspeech', pp. 1883–1887.

El Zarka, D., Schuppler, B. and Cangemi, F. (2019), Acoustic cues to topic and narrow focus in Egyptian Arabic, *in* 'Proceedings of Interspeech 2019', Causal Productions, pp. 1771–1775.

Féry, C. and Kügler, F. (2008), 'Pitch accent scaling on given, new and focused constituents in German', *Journal of Phonetics* **36(4)**, 680–703.

Greif, M. (2010), 'Contrastive Focus in Mandarin Chinese', *In Proceedings of Speech Prosody 2010, Chicago* .

Grice, M., Ladd, D. R. and Arvaniti, A. (2000), 'On the Place of Phrase Accents in Intonationl Phonology', *Phonology* **17**, 143–185.

Gu, C. (2014), 'Smoothing Spline ANOVA Models: R Package gss', *Journal of Statistical Software* **58**(1), 1–25.

Gu, W. and Lee, T. (2007), Effects of focus on prosody of Cantonese speech–A comparison of surface feature analysis and model-based analysis, *in* 'Proceedings of the International Workshop Paralinguistic Speech'07', pp. 59–64.

Harnsberger, J. D. (1994), 'Towards an intonational phonology of Hindi', *In Fifth Conference on Laboratory Phonology. Chicago: Northwestern University* .

Hellmuth, S. (2006*a*), 'Focus-related pitch range manipulation (and peak alignment effects) in Egyptian Arabic', *Proceedings of Speech Prosody 2006* pp. 410–413.

Hellmuth, S. (2007), 'The relationship between prosodic structure and pitch accent distribution: Evidence from Egyptian Arabic', *The Linguistic Review* **24**(2-3), 291–316.

Hellmuth, S. (2009), The (absence of) prosodic reflexes of given/new information status in Egyptian Arabic, *in* J. Owens and A. Elgibali, eds, 'Information structure in spoken Arabic', pp. 165–188.

Hellmuth, S. (2011), 'Acoustic cues to focus and givenness in Egyptian Arabic', *Instrumental studies in Arabic phonetics* **319**, 301.

Hellmuth, S. (2018), Variation in polar interrogative contours within and between Arabic dialects, *in* 'Speech Prosody 2018', ISCA, pp. 989–993.

Hellmuth, S. J. (2006*b*), Intonational pitch accent distribution in Egyptian Arabic, PhD thesis, School of Oriental and African Studies, University of London.

Himmelmann, N. P. and Ladd, R. (2008), 'Prosodic Description: An Introduction for Fieldworkers', *Language Documentation & Conservation* **2**, 244–274.

Holes, C. (1990), *Gulf Arabic*, London: Croom Helm.

Hu, N., Janssen, B., Hansen, J., Gussenhoven, C. and Chen, A. J. (2020), 'Automatic analysis of speech prosody in Dutch', *Paper presented at the Interspeech, October 25–29, 2020, Shanghai, China. .*

Ibrahim, O., El-Ramly, S. and Abdel-Kader, N. (2001), A model of F0 contour for Arabic affirmative and interrogative sentences, *in* 'Proceedings of the Eighteenth National Radio Science Conference. NRSC'2001 (IEEE Cat. No.01EX462)', Vol. 2, Acad. Sci. Res. & Technol, Mansoura, Egypt, pp. 517–524.

Ipek, C. (2011), Phonetic Realization of Focus with No On-focus Pitch Range Expansion in Turkish., *in* 'ICPhS', pp. 140–143.

Ishihara, S. (2002), 'Syntax-Prosody Interface of Wh-Constrcutions in Japanese', *Proceedings of Tokyo Conference on Psycholinguistics (TCP 2002)* pp. 165–189.

Ishihara, S. (2003), Intonation and interface conditions, PhD thesis, Massachusetts Institute of Technology.

Ishihara, S. (2016), 'Japanese downstep revisited', *Natural language & linguistic theory* **34**(4), 1389–1443.

Kirkpatrick, S., Gelatt Jr, C. D. and Vecchi, M. P. (1983), 'Optimization by simulated annealing', *Science (New York, N.Y.)* **220**(4598), 671–680. Publisher: American Association for the Advancement of Science.

Kiss, K. (1998), 'Identificational focus versus information focus', *Language* **74**(2), 245–273.

Krifka, M. (2008), 'Basic notions of information structure', *Acta Linguistica Hungarica* **55**, 243–276.

Kügler, F. and Féry, C. (2017), 'Postfocal downstep in German', *Language and speech* **60**(2), 260–288.

Kügler, F. and Genzel, S. (2012), 'On the prosodic expression of pragmatic prominence: The case of pitch register lowering in Akan', *Language and speech* **55**(3), 331–359.

Ladd, D. R. (2008), *Intonational Phonology*, Cambridge, UK: Cambridge University Press.

Lee, A. and Xu, Y. (2012), Revisiting focus prosody in Japanese, *in* 'Speech Prosody 2012'.

Lee, A., Xu, Y. and Prom-on, S. (2014), Modeling Japanese F0 contours using the PENTA-trainers and AMtrainer, Nijmegen, p. 4.

Lee, Y.-c. and Xu, Y. (2010), Phonetic realization of contrastive focus in Korean, *in* 'Speech Prosody 2010-Fifth International Conference'.

Leung, T. (2014), 'The preposition stranding generalization and conditions on sluicing: Evidence from Emirati Arabic', *Linguistic Inquiry* **45**(2), 332–340.

Leung, T. T.-C., Ntelitheos, D. and Al Kaabi, M. (2021), *Emirati Arabic: A Comprehensive Grammar*, Routledge.

Liu, F., Prom-on, S., Xu, Y. and Whalen, D. H. (2015), 'Computational modelling of double focus in American English'.

Liu, F. and Xu, Y. (2005), ‘Parallel encoding of focus and interrogative meaning in Mandarin intonation’, *Phonetica* **62**, 70–87.

Liu, F., Xu, Y., Prom-on, S. and Yu, A. C. L. (2013), ‘Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling’, *Journal of Speech Sciences* **3(1)**, 85–140.

Liu, X., Xu, Y., Zhang, W. and Tian, X. (2021), ‘Multiple prosodic meanings are conveyed through separate pitch ranges: Evidence from perception of focus and surprise in mandarin chinese’, *Cognitive, Affective, & Behavioral Neuroscience* **21**(6), 1164–1175.

Mahanta, S., Das, K. and Gope, A. (2016), On the Phonetics and Phonology of Focus marking in Boro, *in* ‘Proceedings of the Annual Meetings on Phonology’, Vol. 3.

Mahfoudhi, A. (2002), ‘Agreement lost, agreement regained: A minimalist account of word order and agreement variation in Arabic’, *California Linguistic Notes* **27**(2), 1–28.

Mixdorff, H. (2004), ‘Quantitative Tone and Intonation Modeling across Languages’, *Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages. Beijing* pp. 137–142.

Molnár, V. (2002), Contrast–from a contrastive perspective, *in* ‘Information structure in a cross-linguistic perspective’, Brill Rodopi, pp. 147–161.

Molnár, V. and Winkler, S. (2010), ‘Edges and gaps: Contrast at the interfaces’, *Lingua. International review of general linguistics. Revue internationale de linguistique générale* **120**, 1392–1415.

Moutaouakil, A. (1989), *Pragmatic functions in a functional grammar of Arabic*, Dordrecht: Foris.

Nagahara, H. (1994), Phonological Phrasing in Japanese, PhD Thesis, PhD dissertation, University of California.

Norlin, K. (1989), ‘A preliminary descriptions of Cairo Arabic intonation of statements and questions’, *Speech Transmission Quarterly Progress and Status Report* **1**, 47–49.

Ouhalla, J. (1997), ‘Remarks on focus in Standard Arabic’, *In M. Eid and R. R. Ratcliffe (Eds.), ‘Perspectives on Arabic Linguistics X: Papers from the Tenth Annual Symposium on Arabic Linguistics’* pp. 9–45.

Owens, J., Dodsworth, R. and Kohn, M. (2013), ‘Subject expression and discourse embeddedness in Emirati Arabic’, *Language Variation and Change* **25**, 255–285.

Pan, H.-H. (2008), Focus and Taiwanese unchecked tones, *in* ‘Topic and Focus’, Springer, pp. 195–213.

Patil, U., Kentner, G., Gollrad, A., Kügler, F., Féry, C. and Vasishth, S. (2008), ‘Focus, word order and intonation in Hindi’, *Journal of South Asian Linguistics* **1**.

Pell, M. D. (2001), ‘Influence of emotion and focus on prosody in matched statements and questions’, *Journal of the Acoustical Society of America* **109**, 1668–1680.

Pierrehumbert, J. (1980), The Phonology and Phonetics of English Intonation, PhD Thesis, Bloomington, Ind.: Indiana University Linguistics Club.

Pierrehumbert, J. (1981), 'Synthesizing intonation', *The Journal of the Acoustical Society of America* **70**(4), 985–995.

Pierrehumbert, J. and Beckman, M. E. (1988), *Japanese Tone Structure*, Cambridge, MA: MIT Press.

Prom-on, S., Liu, F. and Xu, Y. (2012), 'Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling', *The Journal of the Acoustical Society of America* **132**(1), 421–432.

Prom-on, S., Xu, Y. and Thipakorn, B. (2009), 'Modeling tone and intonation in Mandarin and English as a process of target approximation', *Journal of the Acoustical Society of America* **125**, 405–424.

R Core Team (2019), R: A language and environment for statistical computing, Manual, Vienna, Austria.

Raidt, S., Bailly, G., Holm, B. and Mixdorff, H. (2004), Automatic generation of prosody: Comparing two superpositional systems, *in* 'Speech prosody 2004, international conference'.

Rialland, A. and Robert, S. (2001), 'The intonational system of Wolof', **39**(5), 893–939.

Rifaat, K. (2005), 'The Structure of Arabic Intonation: A Preliminary Investigation', *In M. T. Alhawary and E. Benmamoun (eds.),'Perspectives on Arabic Linguistics XVII-XVIII. Papers from the Seventeenth and Eighteenth Annual Symposia on Arabic Linguistics'* pp. 49–69.

Rooth, M. (1985), Association with focus, PhD thesis, University of Massachusetts, Amherst.

Rooth, M. (1992), 'A theory of focus interpretation', *Natural Language Semantics* **1**(1), 75–116.

Rosenberg, A. (2010), 'AuToBI – A Tool for Automatic ToBI annotation', *Paper presented at the Eleventh Annual Conference of the International Speech Communication Association (Interspeech).* p. 5.

Rump, H. H. and Collier, R. (1996), 'Focus Conditions and Prominence of Pitch–Accented Syllables', *Language and Speech* **39 (1)**, 1–17.

Sahkai, H., Kalvik, M.-L. and Mihkla, M. (2013), Prosody of contrastive focus in Estonian., *in* 'in Proceedings of Interspeech 2013, Lyon, France, 315-319', pp. 315–319.

Sakurai, A., Hirose, K. and Minematsu, N. (2003), 'Data-driven generation of F0 contours using a superpositional model', *Speech Communication* **40**(4), 535–549.

Searle, S. R., Speed, F. M. and Milliken, G. A. (1980), 'Population marginal means in the linear model: an alternative to least squares means', *The American Statistician* **34**(4), 216–221.

Shlonsky, U. (1997), *Clause structure and word order in Hebrew and Arabic: An essay in comparative Semitic syntax*, New York/ Oxford: Oxford University Press.

Sityaev, D. and House, J. (1819), Phonetic and phonological correlates of broad, narrow and contrastive focus in English, *in* 'Proc. of the 15th ICPhS, Barcelona', Vol. 1822.

Sugahara, M. (2003), Downtrends and Post-focus Intonation in Tokyo Japanese, PhD thesis, University of Massachusetts Amherst.

Sun, X. and Xu, Y. (2002), 'Perceived pitch of synthesized voice with alternate cycles', *Journal of Voice* **16**(4), 443–459.

Szreder, M. and Ben-Ammar, C. (n.d.), Affricate variation in Emirati Arabic: An exploratory study, *in* D. Ntelitheos and T. Leung, eds, 'Experimental Arabic Linguistics', John Benjamins.

Szreder, M., De Ruiter, L. E. and Ntelitheos, D. (2021), 'Input effects in the acquisition of verb inflection: Evidence from Emirati Arabic.', *Journal of Child Language* .

Team, A. (2020), 'Audacity(R): Free Audio Editor and Recorder [Computer application]'.
URL: *https://audacityteam.org*

Truckenbrodt, H. (1995), Phonological phrases: Their relation to syntax, focus and prominence, PhD Thesis, Linguistics, MIT.

Uechi, A. (1998), An interface approach to Topic/Focus structure, PhD thesis, University of British Columbia.

Vallduví, E. and Vilkuna, M. (1998), On rheme and kontrast, *in* P. Culicover and L. McNally, eds, 'The limits of syntax', Brill, pp. 79–108.

Wang, B., Qadir, T. and Xu, Y. (2013), 'Prosodic encoding and perception of focus in Uygur', *Chinese Journal of Acoustics. in press.(in Chinese)* .

Wang, B. and Xu, Y. (2011), 'Differential prosodic encoding of topic and focus in sentence-initial position in Mandarin Chinese', *Journal of Phonetics* **39**, 595–611.

Wang, B., Xu, Y. and Ding, Q. (2018), 'Interactive prosodic marking of focus, boundary and newness in Mandarin', *Phonetica* **75**(1), 24–56.

Wang, L., Wang, B. and Xu, Y. (2012), 'Prosodic encoding and perception of focus in Tibetan (Anduo Dialect)', *Speech Prosody 2012. Shanghai* pp. 286–289.

Watson, J. C. (2002), *The phonology and morphology of Arabic*, Oxford: Oxford University Press.

Wenjun, D. and Yuan, J. (2015), Contrastive study of focus phonetic realization between Jinan dialect and Taiyuan dialect, *in* '2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)', IEEE, Shanghai, China, pp. 47–52.

Woods, K. J. P., Siegel, M. H., Traer, J. and McDermott, J. H. (2017), 'Headphone screening to facilitate web-based auditory experiments', *Attention, Perception & Psychophysics* **79**(7), 2064–2072.

Xu, Y. (1998), 'Consistency of tone-syllable alignment across different syllable structures and speaking rates', *Phonetica* **55**(4), 179–203.

Xu, Y. (1999), 'Effects of tone and focus on the formation and alignment of F0 contours', *Journal of Phonetics* **27**(1), 55–105.

Xu, Y. (2005), 'Speech melody as articulatorily implemented communicative functions', *Speech communication* **46**(3-4), 220–251.

Xu, Y. (2010), 'In defense of lab speech', *Journal of Phonetics* **38**(3), 329–336.

Xu, Y. (2011), Post-Focus Compression: Cross-Linguistic Distribution and Historical Origin, *in* 'Proceedings of The 17th International Congress of Phonetic Sciences, Hong Kong'.

Xu, Y. (2013), 'ProsodyPro – a tool for large-scale systematic prosody analysis', *TRASP* pp. 7–10.

Xu, Y., Chen, S.-w. and Wang, B. (2012), 'Prosodic focus with and without post-focus compression: A typological divide within the same language family?', *The Linguistic Review* **29**(1), 131–147.

Xu, Y., Lee, A., Prom-on, S. and Liu, F. (2015), 'Explaining the PENTA model: A reply to Arvaniti and Ladd', *Phonology* **32**, 505–535.

Xu, Y. and Liu, F. (2012), 'Intrinsic coherence of prosodic and segmental aspects of speech', *Understanding prosody: The role of context, function and communication. Walter de Gruyter* pp. 1–26.

Xu, Y. and Prom-on, S. (2014), 'Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via Model–Based stochastic learning', *Speech Communication* **57**, 181–208.

Xu, Y., Prom-on, S. and Liu, F. (2022), 'The PENTA model: Concepts, use and implications', *In Prosodic Theory and Practice (S. Shattuck-Hufnagel & J. Barnes, editors)* pp. 377–407. Cambridge: The MIT Press.

Xu, Y. and Wang, M. (2009), 'Organizing syllables into groups_evidence from F0 and duration patterns in Mandarin', *Journal of phonetics* **37**(4), 502–520.

Xu, Y. and Wang, Q. E. (2001), 'Pitch targets and their realization: Evidence from Mandarin Chinese', *Speech communication* **33**(4), 319–337.

Xu, Y. and Xu, C. X. (2005), 'Phonetic realization of focus in English declarative intonation', *Journal of Phonetics* **33**(2), 159–197.

Yeou, M., Embarki, M. and Al-Maqtari, S. (2007), 'Contrastive focus and F0 patterns in three Arabic dialects', *Nouveaux cahiers de linguistique française* **28**, 317–326.

Zerbian, S., Genzel, S. and Kügler, F. (2010), Experimental work on prosodically-marked information structure in selected African languages (Afroasiatic and Niger-Congo), *in* 'Speech Prosody 2010-Fifth International Conference'.

Zimmermann, M. (2008), 'Contrastive focus and emphasis', *Acta Linguistica Hungarica* **55**, 347–360.