COMPUTATIONAL MODELLING OF DOUBLE FOCUS IN AMERICAN ENGLISH

Fang Liu[1], Yi Xu[1], Santitham Prom-on[1,2] & D. H. Whalen[3]

[1]Department of Speech, Hearing and Phonetic Sciences, University College London, UK;
[2]Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand;
[3]City University of New York, New York, NY, and Haskins Laboratories, New Haven, CT, USA.
liufang@uchicago.edu; yi.xu@ucl.ac.uk; santitham@cpe.kmutt.ac.th; whalen@haskins.yale.edu

ABSTRACT

This study investigated how double focus in English statements and questions can be computationally modelled. PENTAtrainer2 was used to learn syllable-sized multi-functional targets from a corpus of 1960 English utterances, with controlled variations in lexical stress, focus, modality, and sentence length. The results showed that the learned targets could generate F_0 contours close to the original. In particular, the asymmetry in the interaction between focus and modality was effectively simulated.

Keywords: Focus, modality, English, pitch target, PENTAtrainer.

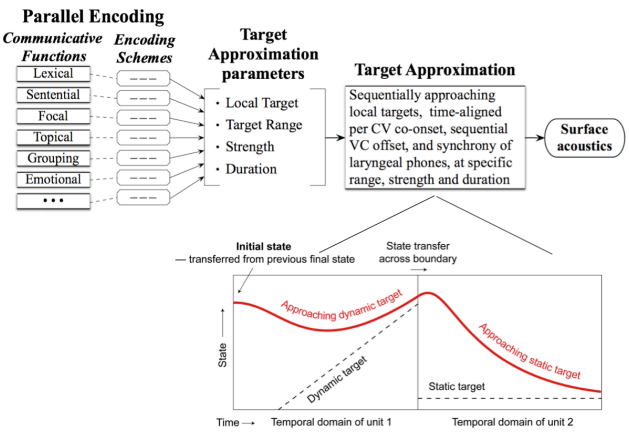
1. INTRODUCTION

In spoken languages, focus prosodically highlights specific contents in an utterance [4]. Typically, only one component of an utterance needs to be highlighted, and such single focus has been studied for a variety of languages in both statements and questions, as reviewed in [5, 31]. In certain situations, however, more than one element in an utterance needs to be focused [18]. There have been a limited number of studies on the prosody of multiple foci, including American English [9], Dutch [25], German [28], and Mandarin Chinese [15, 28]. These studies investigated focus only in statements, however.

A preliminary study on multiple foci in questions in English has established some basic patterns [19]. In general, single and double focus are realized similarly in statements, but differ in terms of post-focus pitch modification in yes/no questions. In statements, double focus has similar effects as single focus on maximum F_0 and duration in statements. In yes/no questions, however, there is post-focus F_0 lowering in double focus, which differs from single focus, where there was no post-focus F_0 lowering. This is likely due to the fact that focus and modality modify F_0 of the between-focus words in opposite directions in double-focused yes/no questions. This finding adds to previous findings that communicative functions interact with one another in complex ways, so that any particular function, such as focus or modality (sentence type), cannot be said to have a single set of features that are fully autonomous from other functions [21].

These findings pose challenges to computational modelling of prosody, especially to models that treat communicative functions as the building blocks of prosody [1, 11, 30]. This is because, if functional components are combined through superposition, the functional interaction as seen in the case of double focus cannot be directly handled. What is needed is a way of accommodating full and free interaction of multiple functions. In this study, we tested a computational approach based on the PENTA model of prosody [30]. As shown in Fig. 1, PENTA assumes that the syllable is the basic prosody carrier, whose duration, intensity, pitch target and phonation register are the building blocks of prosody. Multiple communicative functions jointly determine the parameters of the building blocks in each syllable, which in turn control the articulatory process that generates surface prosody. PENTAtrainer2, a program that uses machine learning algorithms to automatically extract parameters of pitch targets in PENTA from functionally annotated speech corpora, has been developed to test this model [32].

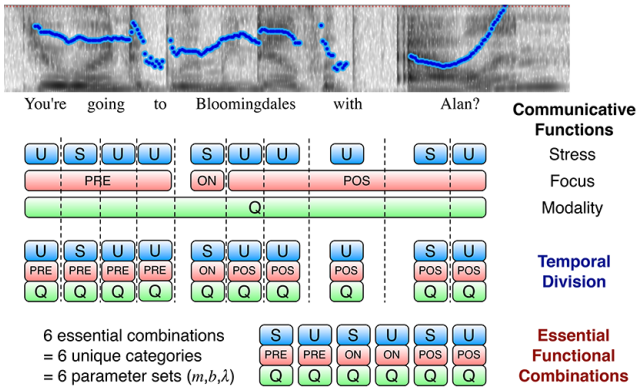
Figure 1: A sketch of the Parallel Encoding and Target Approximation (PENTA) model [30].



In this approach, the accommodation of functional interaction is implemented with three strategies: (a) layered annotation, (b) pseudo-hierarchical representation, and (c) edge-synchronization, as illustrated in Fig. 2. Layered annotation allows clear

separation of individual communicative functions. Pseudo-hierarchical representation means that functional layers are arbitrarily ordered, so that no function dominates others. But because functions differ in their global domain as well as optional internal subdomains (e.g., focus is divided into pre-focus, on-focus and post-focus intervals [7, 8, 22, 33]), when they interact with each other, their domains and subdomains are combined, and the layer with the smallest temporal units projects to layers with larger intervals. Finally, edge-synchronization ensures that all the layers, regardless of their own temporal scope, have fully synchronized edges with the smallest unit, which is the syllable. As a result, each syllable has to bear a combination of the effects of all the functions present in the utterance. An important consequence of such functional combination is that, for each syllable in a particular utterance, only a single multi-functional target is needed.

Figure 2: Illustration how PENTAtrainer2 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 question. From [32].



The single multi-functional target of each syllable is specified in terms of its height, slope, and rate of approximation in the Target Approximation model in PENTA, as depicted in the bottom panel of Fig. 1. These pitch targets can be computationally realized through the qTA model [24]:

$$(1) \qquad f_0(t) = (mt + b) + (c_1 + c_2 t + c_3 t^2)\, e^{-\lambda t},$$

where $f_0(t)$ is the F_0 value at time t in semitone scale (st), t is the relative time (in s) from the onset of the syllable, m is the slope of the target (in st/s), b is the height of the target (in st) defined as the intercept of the target offset with the y-axis, λ is the strength of the target approximation movement, and c_i are coefficients to be derived from the initial state and the target using the following formulae.

$$(2) \qquad c_1 = f_0(0) - b$$
$$(3) \qquad c_2 = f_0'(0) + c_1\lambda - m$$
$$(4) \qquad c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2$$

Given that qTA can generate continuous F_0 contours with a sequence of syllable-sized pitch targets, it is possible to optimize the targets through analysis-by-synthesis [11, 14, 26]. That is, candidate targets are modified iteratively to minimize the error between synthetic F_0 trajectories and those of natural speech. This optimization process stops once the error converges to a very small error threshold or the maximum number of steps is reached. This process is implemented in PENTAtrainer2, which optimizes for the multi-functional targets with three tools: Annotation, Learning, and Synthesis [32]. The Annotation tool enables the annotation of multiple functional layers. The Learning tool optimizes qTA parameters by *simulated annealing*, a machine learning algorithm that globally explores the target search space and iteratively locates a good approximation of the global optimum solution. The Synthesis tool generates continuous F_0 contours based on targets that have been optimized for all the functional combinations, and provides direct graphical, numerical and auditory comparisons between the original and synthesized prosody [32].

PENTAtrainer2 has been used to model single focus in English with promising results [32]. In particular, it was shown that the multi-functional target optimization approach was able to simulate the asymmetrical interactions of lexical stress, focus and modality in English. The current investigation aims to apply a similar strategy to model double focus in English, with the goal of further assessing PENTAtrainer2's ability to process the interactions of not only focus and modality, but also a number of other functions that are frequently present in any speech utterance.

2. METHOD

2.1. Corpus

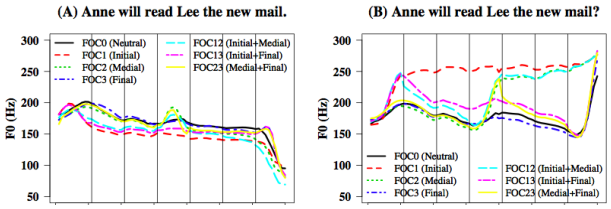
The speech materials contained four sets of English statements and yes/no questions of different lengths (short, medium, long, extra long), in which the initial, medial, and final words (italicized) were designated as focus-bearing key words.

1. **Short (7 syll.)**: *Anne* will read *Lee* the new *mail*
2. **Medium (12 syllables)**: *Nina* is selling *Lily* a yellow *lemon*
3. **Long (17 syll.)**: *Elaine* might be introducing *Lamar* to her best girlfriend *Arlene*

4. **Extra long (24 syll.)**: *Amelia* has been accommodating *Ramona* with a lot of delicious *vanilla*

Each of the four utterance sets consisted of fourteen distinct utterances with seven focus (neutral, initial, medial, final, initial + medial, initial + final, medial + final) and two modality conditions (statement, yes/no question), prompted by different priming statements/questions. Fig. 3 shows mean time-normalized F_0 contours of the first set of utterances. The other three sets show similar patterns [19]. Each utterance was produced five times by every subject, resulting in 1960 utterances in total (4 lengths × 2 modalities × 7 focus conditions × 7 subjects × 5 repetitions).

Figure 3: Time-normalized F_0 contours of the first set of utterances, averaged across 35 tokens (5 repetitions × 7 subjects), data from [19]. Note: FOC0 (Neutral): an utterance with neutral focus; FOC12 (Initial + Medial): an utterance with double focus on the initial and medial key words. The vertical lines denote syllable boundaries.



2.2. Modelling

Though originally designed to study the interaction of focus and modality, the sentences in the corpus also naturally carry a number of other functions, including lexical stress, syllable position and part of speech. Giving them explicit representations would allow us to assess their specific contributions to prosody, and could also potentially enhance the quality of modelling. The data were therefore annotated in terms of the following functions after initial processing with ProsodyPro [29]:

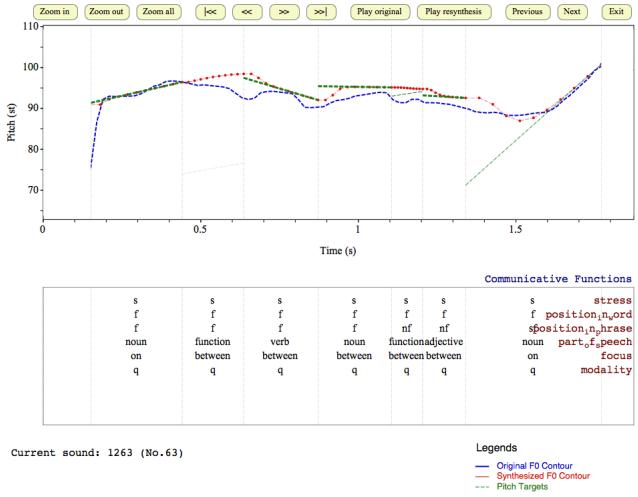
1. **Stress of syllable**: unstressed, stressed
2. **Focus**: pre-focus, on-focus, post-focus, between-focus
3. **Modality**: statement, question
4. **Syllable position in phrase**: non-final, sentence-final, final
5. **Syllable position in word**: non-final, final
6. **Part of speech**: noun, function, verb, adjective

To compare the contribution of each of the six functions, for each individual speaker, learning and synthesis were done first with all six functional layers included, and then with each one of the functions removed. Cross-speaker learning and synthesis of the whole set of 1960 utterances were also tested with all six functional layers included. With all 6 functions included, there were 146 unique multi-functional targets to be learned from this corpus. Excluding each of the 6 functions resulted in different reductions of the total number of targets. In the Learning phase, the optimization parameters of PENTAtrainer2 were set as follows: Maximum Iteration = 500, Learning Rate = 0.1, Starting Temperature = 700, and Reduction Factor = 0.9.

3. RESULTS

Following [32], the accuracy of synthesis was evaluated by root-mean-square error (RMSE), which indicates the average mismatch between the synthetic and original pitch contours, and Pearson's correlation (correlation hereafter), which indicates the strength and direction of the linear relationship between the synthetic and original pitch contours [13]. Fig. 4 shows an example of the original and synthetic pitch contours of an utterance, with annotations of the six functions.

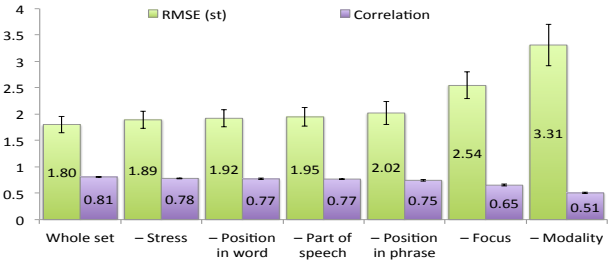
Figure 4: The demo window of the synthesis tool of PENTAtrainer2, showing the original contours of an utterance by a female speaker (dotted blue curve), the learned pitch targets (green dashed lines), and the synthetic contours (red dotted curve). The lower panel shows the six functional layers, with annotations for every syllable.



The cross-speaker synthesis on the whole set of 1960 utterances with all six functions included yielded an RMSE of 2.750 st (SD = 0.685), and a correlation of 0.687 (SD = 0.051). The averaged synthesis accuracy based on each individual speaker's synthesis data with all six functions included achieved better performance than across-speaker synthesis: RMSE = 1.804 st (SD = 0.417), and correlation = 0.809 (SD = 0.025). This improvement is consistent with previous findings [24, 32].

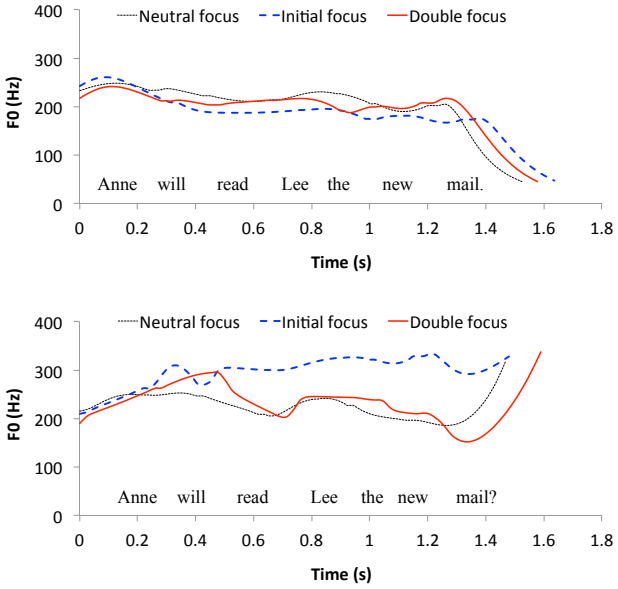
Fig. 5 shows detailed accuracy results averaged across individual speakers' synthesis data. As can be seen, the importance of each of the six linguistic functions showed the rank of modality > focus > syllable position in phrase > part of speech > syllable position in word > stress, as demonstrated by the amount of accuracy reduction (larger RMSE, smaller correlation) from the synthesis based on the whole set of functions. Repeated measures ANOVAs revealed a significant effect of function on both RMSE ($F(6, 36) = 28.4$, $p < 0.001$) and correlation ($F(6, 36) = 165.8$, $p < 0.001$). For RMSE, post-hoc pairwise t tests with the Holm correction showed that there were significant differences between the "–modality" condition and all other conditions (all ps < 0.01), but no other differences were significant (all ps > 0.05). For the correlations, significant differences were found between "–modality" and other conditions (all ps < 0.001), between "–focus" and other conditions (all ps < 0.001), and between "the whole set" and "–position in phrase" ($p = 0.007$). No other differences were significant (all ps > 0.05).

Figure 5: Mean RMSEs (in st) and correlation coefficients of the synthesized F_0 against the original, with different functional combinations, averaged across 7 speakers. "Whole set" means all 6 functions were included in learning and synthesis, and "–Stress" means the stress function was excluded, and so on. Error bars represent standard errors.



These results show that each of the tested functions made certain contributions to the F_0 contours of these sentences. But by far the most important functions are focus and modality. The importance of these two functions is further illustrated in Fig. 6, which shows examples of synthesized pitch contours of the sentence "Anne will read Lee the new mail" as either a statement (upper panel) or a question (lower panel), and with different types of focus (within each panel) modelled on one of the female speakers. The pitch target parameters used in synthesis were trained on and applied to the data of the same speaker. The effect of focus can be seen in the differences between the three synthetic F_0 tracks in both panels. But the differences are much larger in the lower panel than in the upper panel, which are similar (though not identical, due to cross-speaker differences) to the natural contours in Fig. 3.

Figure 6: Synthetic F_0 contours generated with pitch targets learned from the speech of one of the female speakers.



4. DISCUSSION AND CONCLUSION

The results this study demonstrate that asymmetries in functional interactions such as those between focus and modality can be handled with the modelling approach implemented in PENTAtrainer2. That is, by treating each syllable as a carrier of multiple functions, unique syllable-sized targets can be learned from a functionally annotated corpus. These targets can then be applied in synthesis to syllables with identical functional combinations to generate naturalistic F_0 contours. Thus this approach can simultaneously accommodate two conventional theories about the nature of intonation that sharply oppose each other, namely, the linear [2, 23] and superpositional [1, 11, 12] views. Linearity is achieved in our approach by requiring only a single sequence of pitch targets to generate surface prosody, no matter how many functions are involved. Superposition is encompassed in the multi-functionality of every syllable-sized target. Thus the approach has found a middle way, as has been hoped [17], between the two extreme positions. But a further question is whether the approach in any way reflects reality. For example, is it really likely that speakers learn and apply so many multi-functional targets syllable by syllable, and listeners decode the functions sequentially rather than separately processing local contours and global profiles? This kind of question calls for further behavioural as well as neural studies that examine the importance of sequential information processing in both production and perception [6, 27].

5. REFERENCES

[1] Bailly, G., Holm, B., 2005. SFC: a trainable prosodic model. *Speech Communication* 46: 348-364.

[2] Beckman, M. E., 1995. *Local shapes and global trends*. *Proc. 13th ICPhS*, Stockholm, 100–107.

[3] Boersma, P. and Weenink, D. 2001. Praat, a system for doing phonetics by computer. *Glot international*. 5, 341–345.

[4] Bolinger, D. 1972. Accent Is Predictable (If You're a Mind-Reader). *Language*. 48, 633–644.

[5] Breen, M., Fedorenko, E., Wagner, M. and Gibson, E. 2010. Acoustic correlates of information structure. *Language and Cognitive Processes*. 25, 1044–1098.

[6] Conway, C. M., Pisoni, D. B., Kronenberger, W. G., 2009. The Importance of Sound for Cognitive Sequencing Abilities: The Auditory Scaffolding Hypothesis. *Current Directions in Psychological Science* 18: 275-279.

[7] 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.

[8] Eady, S.J. and Cooper, W.E. 1986. Speech intonation and focus location in matched statements and questions. *The Journal of the Acoustical Society of America*. 80, 402–415.

[9] 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, 233–251.

[10] Fletcher, J. and Evans, N. 2002. An acoustic phonetic analysis of intonational prominence in two Australian languages. *Journal of the International Phonetic Association*. 32, 123–140.

[11] Fujisaki, H., Wang, C., Ohno, S., Gu, W., 2005. Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command–response model. *Speech communication* 47: 59-70.

[12] Grønnum, N., 1995. *Superposition and subordination in* intonation — *a non-linear approach*. *Proc. 13th ICPhS*, Stockholm,124-131.

[13] Hermes, D.J. 1998. Measuring the perceptual similarity of pitch contours. *Journal of speech, language, and hearing research: JSLHR*. 41, 73–82.

[14] Hirst, D., 2011. The analysis by synthesis of speech melody: From data to models. *Journal of Speech Sciences* 1: 55-83.

[15] Kabagema-Bilan, E., López-Jiménez, B. and Truckenbrodt, H. 2011. Multiple focus in Mandarin Chinese. *Lingua*. 121, 1890–1905.

[16] Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. 1983. Optimization by Simulated Annealing. *Science*. 220, 671–680.

[17] Ladd, D. R., 1995. *"Linear" and "overlay" descriptions: An autosegmental-metrical middle way*. *Proc. 13th ICPhS*, Stockholm, 116-123.

[18] Ladd, D.R. 2009. *Intonational Phonology*. Cambridge University Press.

[19] Liu, F. 2010. Single vs. double focus in English statements and yes/no questions. *Proceedings of Speech Prosody 2010* (Chicago, USA, 2010).

[20] Liu, F. and Xu, Y. 2005. Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica*. 62, 70–87.

[21] 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, 85–140.

[22] Pell, M.D. 2001. Influence of emotion and focus location on prosody in matched statements and questions. *The Journal of the Acoustical Society of America*. 109, 1668–1680.

[23] Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation.* Ph.D. dissertation. MIT, Cambridge, MA.

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

[25] Rump, H.H. and Collier, R. 1996. Focus conditions and the prominence of pitch-accented syllables. *Language and Speech*. 39, 1–17.

[26] Stevens, K. N., Halle, M., 1967. Remarks on analysis by synthesis and distinctive features. In *Models for the Perception of Speech and Visual Form*. W. Wathen-Dunn. MIT, Cambridge, MA**:** 88-102.

[27] van Berkum, J. J. A., Zwitserlood, P., Hagoort, P., Brown, C. M., 2003. When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research* 17, 701-718.

[28] Wang, B. 2008. The prosodic realization of dual focus: Comparing Mandarin and German. *Proceedings of the 8th Phonetic Conference of China and the International Symposium on Phonetic Frontiers* (2008).

[29] Xu, Y. 2013. ProsodyPro — A tool for large-scale systematic prosody analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody* (Aix-en-Provence, France, 2013), 7–10.

[30] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication*. 46, 220–251.

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

[32] 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.

[33] Xu, Y. and Xu, C.X. 2005. Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*. 33, 159–197.