TOWARD PREDICTIVE MODELLING FOR AM THEORY OF INTONATION

Emily Lau, Yi Xu
University College London
emily.nh.lau91@gmail.com, yi.xu@ucl.ac.uk

ABSTRACT

This paper presents the first version of AMtrainer that can computationally parameterize tonal categories of the AM Theory of intonation. The parameterization was done by specifying each tonal category as the mean F_0 of all the individual tokens of the category in a corpus. We trained AMtrainer on a 192-sentence corpus in American English and used the learned parameters to synthesize the prosody of those sentences. The numerical accuracy of the synthetic prosody and listener judgments of focus, sentence type and naturalness indicate that the learned categorical parameters can generate fairly natural F_0 contours that convey pragmatic meanings such as sentence type and focus. More importantly, this new development in AMtrainer makes AM theory more directly comparable than before to other computationally implemented theories of prosody.

Keywords: AM Theory, AMtrainer, prosody, predictive synthesis, global synthesis.

1. INTRODUCTION

Technological advances have prompted a movement to computationally implement prosodic theories to the extent that they can automatically learn numerical parameters from speech data and apply the learned parameters to generate intonational configurations for novel utterances [5, 7-8, 13, 20]. The Common Prosody Platform (CPP) was developed to facilitate this movement by making various prosodic theories directly comparable to each other in terms of their capacity to predict prosodic configurations [16]. Among the theories implemented was the well-known Autosegmental-Metrical (AM) theory [3, 14], in the form of the computational program AMtrainer [1]. However, this and another early version of AMtrainer [11] were only able to perform direct fitting of local F_0 contours, so that different instances of a tone category have different F_0 values rather than sharing a category-specific common value. Such a model therefore still lacks predictive power.

There have also been other computational models of intonation that implement the concepts of the AM theory. But they vary to the extent that they adhere to the core notion that intonation consists of contrastive tonal targets realized as prominent F_0 turning points connected via target-transition [3-4, 14]. Some have implemented this core notion literally [15], while others have added additional algorithms such as smoothing filters [2]. There have also been wide uses of the ToBI system of intonation annotation [17] in which the tonal elements are derived from AM theory. However, none of those applications has implemented the theory's core concept of target-transition. Most critically, there have been a lack of open-access systems that allow direct quantitative comparison of AM theory to other theories.

The aims of the present study are to **1)** develop an enhanced version of AMtrainer with the capacity to learn numerical parameters of tone categories of AM theory that can be used to generate continuous F_0 configurations for sentences that have been tonally annotated, and **2)** use the results of this updated AMtrainer to test the validity of the fundamental assumptions of AM Theory.

2. METHODOLOGY

2.1 Developing AMtrainer

The AM theory posits that English intonation consists of tones that are either H (high) or L (low) that are manifested as turning points in surface F_0 contours, and transitions between them that are either linear or curvilinear [3, 14]. This scheme was faithfully implemented in a previous version of AMtrainer [11]. In that version, however, there was no categorization of, or ways to empirically define the tones, because each F_0 target, including both its height and alignment, remains the same as in the original speech sample, and are directly taken from the annotated targets even during synthesis. This makes that version of AMtrainer largely a curve fitting program. To make the program predictive, it is necessary to achieve categorical parameterization of the tones.

At least two aspects of a tone can be parameterized: F_0 and temporal alignment. Of the two, the first is relatively well defined in the theory, i.e., it specifies both the tone category itself and the pitch range determined by either a baseline or a reference line [2-3]. For alignment, however, there are no clear theory-motivated specifications. There have been reports from empirical research on consistent alignment of F_0 turning points for some

tonal accents [4, 9-10], but so far there is no consensus.

For the *present* study, we tested parameterization of only tone height; and even for this we tested only that of tone-specific target height, without implementing a baseline or reference line [2]. One motivation for this choice is that our prior modelling with PENTAtrainer did not show clear detrimental effects of lack of explicit declination [20].

The parameterization of tone-specific target height is straightforward. We define the height of a tone as the mean F_0 of all the individual tokens of the same tone category in a corpus. In terms of the inventory of tone categories, we considered only those for which there is a clear consensus. In particular, composite tones like LH* and HL* were not included, due to difficulty of distinguishing them from singleton tones. Also not included were L- and H-, whose implementational definitions are not clear. The F_0 between the tonal targets are either straight lines or parabolic curves, following [15]. Technically, AMtrainer is implemented as a Praat [6] script and is openly accessible at [1].

2.2. The corpus

The corpus consists of a total of 192 utterances produced by a female speaker in the experimental corpus collected in a study of American English intonation [12]. It includes 6 unique sentences, each said in two sentence types (declarative vs. question) and two focus conditions (medial vs. final). As shown below, each target utterance is preceded by a leading sentence (in parentheses) that elicited a specific focus (by boldface) and sentence type (by punctuation) condition. Each sentence was repeated 8 times, resulting in a total of 192 utterances.

1. (Not an internship./?) / (Not La Massage./?)
   You want a **job** with Microsoft./?
   You want a job with **Microsoft**./?
2. (Not an internship./?) / (Not Microsoft./?)
   You want a **job** with La Massage./?
   You want a job with **La Massage**./?
3. (It's not fate./?) / (It's not you./?)
   There's something **unmarriable** about me./?
   There's something unmarriable about **me**./?
4. (It's not fate./?) / (It's not **me**/**you**./?)
   There's something **unmarriable** about **May**./?
   There's something **unmarriable** about **May**./?
5. (It's not Sears./?) / (It's not Elaine./?)
   You're going to **Bloomingdales** with Alan./?)
   You're going to Bloomingdales with **Alan**./?)
6. (It's not Sears./?) / (It's not Alan./?)
   You're going to **Bloomingdales** with Elaine./?)

Each utterance was manually labeled in AMtrainer, under the *Interactive view* task, for 7 tone categories: %, L, L*, L%, H, H* and H%. The % tone is unique to AMtrainer, but its inclusion is obligated by the need for an onset F_0 point that specifies the transition to the very first tonal target. The alignment of each tone with its respective position in each sentence remains the same during resynthesis.

2.3. Model training and synthesis

The training was done under the *Get ensemble files* task in AMtrainer. The procedure *All_AM_means* was called to calculate mean F_0 (in semitones) of each of the tones across the whole corpus based on the annotated tone labels. The resulting tonal parameters are shown in Table 1.

Table 1: AM tonal means obtained by AMtrainer from the present corpus.

Tone Name	Tone Height
%	93.52
L	89.20
L*	91.24
L%	80.19
H	93.73
H*	93.72
H%	101.19

These parameters were then used to generate fresh F_0 contours that were used to resynthesize all the sentences in the corpus. In addition, resynthesis was also done with an early version of AMtrainer that performed only local fitting. An example of the resynthesized F_0 contour is shown in Figure 1.

2.4. Listening experiment

Ten native speakers of English, 7 males and 3 females, all students at University College London (UCL), were recruited as subjects in a listening experiment. They heard 3 versions of each of the 24 sentences: the original recording, prosody generated by the previous version of AMtrainer that generated F_0 via local contour fitting, and prosody generated by the new version. The listening tests were done in a quiet room at UCL, and the presentation of the stimuli was administered via Praat's ExperimentMFC interface. Subjects were shown the question corresponding with the experimental task at hand, along with a set of buttons representing the different choices they had to make.

Only one of the 8 repetitions of each of sentence type and focus combination was used, such that 24 of the original 192 utterances were played in each of the three conditions. Each version of these 24 utterances was replicated 3 times, so that each listener heard a total of 216 randomized utterances.

For sentence type judgment, subjects were asked if the sentence was a statement or a question. For focus judgment, they were asked which word in each utterance, or none of them, was emphasized. For naturalness judgment, Mean Opinion Score (MOS) was used, and subjects were asked to rate how natural they felt each sentence sounded, with 1 being "completely unnatural" and 5 being "very natural."

3. RESULTS

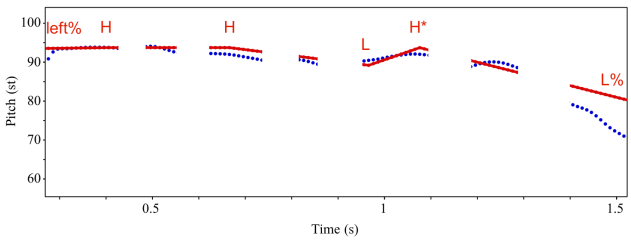
3.1. Synthesis Accuracy

Table 2 shows the RMSE and correlation values for AMtrainer's local fitting and global fitting by stimulus subgroup. Figure 1 shows one example of the globally synthesized F0 curves as compared to the F0 curves of the natural sounds. A repeated measures ANOVA showed that there was significant difference between local and global fitting in both RMSE and correlation (RMSE: $F(1,191) = 11.821$, $p = 0.001$, Partial Eta Squared = 0.058; Correlation: $F(1,191) = 173.302$, $p < 0.001$, Partial Eta Squared = 0.476). From local fitting to global fitting, RMSE increased and correlation decreased, indicating that synthesis accuracy significantly decreased, as expected.

Table 2: Mean RMSE & Pearson correlation for each stimulus subgroup for both local and global synthesis

Sentence Type	Focus	Local fitting		Global fitting	
		RMSE	*r*	RMSE	*r*
Question	Final	1.267	0.901	1.881	0.788
	Medial	1.407	0.941	3.256	0.725
	Sub-avg.	**1.337**	**0.921**	**2.569**	**0.757**
Statement	Final	1.759	0.848	2.8	0.731
	Medial	3.195	0.775	9.724	0.584
	Sub-avg.	**2.477**	**0.812**	**6.262**	**0.658**
Grand average		**2.218**	**0.866**	**4.415**	**0.707**

Figure 1: Globally synthesized F0 curve and natural F0 curve for the sentence "You want a job with Microsoft." Blue dotted line: Natural F0 curve; Red solid line: synthesized F0 curve.



Both types of fittings have significantly decreased accuracy for medial focus, and especially for medial focus statements. However, this decrease in accuracy is much more pronounced in global fitting than in local fitting.

3.2. Sentence type judgment

Table 3 shows percentages of correct judgment of sentence type on stimulus sentences with 3 difference sources of F_0 contours.

Table 3: Percentages of correct responses to the sentence-type judgment by stimulus group

Sentence Type	Focus	Natural	Local fitting	Global fitting	Mean
Question	Final	98.89	98.33	97.22	98.15
	Medial	100	99.44	96.11	98.52
	Sub-perc.	**99.44**	**98.89**	**96.67**	**98.43**
Statement	Final	75	82.22	68.33	75.18
	Medial	88.33	93.89	91.11	91.11
	Sub-perc.	**81.67**	**88.05**	**79.72**	**82.96**
Grand mean		**90.56**	**93.47**	**88.05**	**90.74**

A one-way repeated measures ANOVA showed a significant effect of pitch source ($F(2,18) = 6.733$, $p = 0.007$, Partial Eta Squared = 0.428). The most prominent difference is between local fitting and global fitting ($t(24) = 3.00$, $p = 0.002$), while there were no significant differences between the natural sounds and either synthetic fit ($t(24) = -2.40$, $p = 0.076$; $t(24) = 0.600$, $p = 1.00$). It is also notable that the overall judgments on local fittings were better than those of natural stimuli, which could be explained by some targets being mistakenly given very low F_0 due to inaccuracy of F_0 extraction in Praat, which might have actually assisted listeners' judgments. Regardless of pitch source, judgments for questions were overall more accurate than for statements. It seems that for this corpus, statements are more mistakable for questions. It could be the case that the presence of both medial and final focus has given listeners the impression that the speaker was producing rhetorical questions. It is also possible that since most of the subjects were British English speakers, it was difficult for them to make judgments about American English prosody.

3.3. Focus judgment

The results of focus judgment are shown in Table 4, broken down by subgroups. Errors of judgment were much more common in statements than in questions, and especially when the focus was final. But this drop-off is consistent across natural and synthetic conditions, indicating that focus is generally less detectable in statements.

An ANOVA ($F(2,18)=6.225$, $p = 0.009$, Partial Eta Squared = 0.409) showed that there was a significant difference between natural utterances and global fitting ($t(24) = 3.00$, $p = 0.015$), but not between natural utterances and local fitting ($t(24) = 1.90$, $p = 0.082$). Therefore, even if the effect of the synthetic

conditions is not overwhelming, it cannot be dismissed.

Table 4: Percentages of correct responses to the focus judgment task by stimulus subgroup

Sentence Type	Focus	Original	Local Synthesis	Global Synthesis	Average
Question	Final	95.56	94.44	92.22	94.07
	Medial	95	93.89	89.44	92.78
	Sub-avg	**95.27**	**94.17**	**90.83**	**93.42**
Statement	Final	75.56	71.11	76.67	74.45
	Medial	88.33	82.78	78.89	83.33
	Sub-avg	**81.94**	**76.94**	**77.78**	**78.89**
Grand average		**88.61**	**85.56**	**84.3**	**86.16**

3.4 Naturalness judgment

The naturalness judgment results are shown in Table 5. There is a clear trend of decrease in naturalness from natural to synthetic stimuli, and from local fitting to global fitting. An ANOVA showed a significant effect of pitch source ($F(2,18) = 20.825$, $p < 0.001$, Partial Eta Squared = 0.698).

Table 5: Mean naturalness ratings by stimulus subgroup

Sentence Type	Focus	Original	Local fitting	Global fitting	Average
Question	Final	4.128	3.311	3.022	3.487
	Medial	4.222	3.778	3.022	3.674
	Sub-avg	**4.175**	**3.544**	**3.022**	**3.580**
Statement	Final	4.172	3.694	3.522	3.796
	Medial	4.127	3.339	2.85	3.439
	Sub-avg	**4.15**	**3.517**	**3.186**	**3.618**
Grand average		**4.162**	**3.427**	**3.104**	**3.564**

4. DISCUSSION

The results of both numerical comparisons and perception tests show that the new version of AMtrainer with global fitting based on categorical parameterization of tones can approximate the essential F_0 contours of English intonation that convey focal and sentential meanings. Unsurprisingly, its performance is not as good as the early version of AMtrainer with only local fitting of F_0 contours, which is expected as a model becomes predicative. Also unsurprisingly, even F_0 configurations generated by local fitting were not as good as those of natural speech in terms of both focus and sentence type judgment and naturalness rating. But the development of AMtrainer has at least made such comparisons possible by computationally parameterizing tonal categories of the AM theory.

The development of AMtrainer also makes it possible to compare AM-theory based modeling with models based on other theories of intonation. For example, Table 6 shows mean RMSE and correlation values from [20], which can be compared to the grand averages obtained in the present study shown in Table 2. That study applied PENTAtrainer to the full corpus from which the sub-corpus in the present study was drawn. It used PENTAtrainer to categorically parameterize focus and sentence type through global optimization. In terms of both RMSE and correlation, AMtrainer with local fitting performed largely similar to PENTAtrainer. However, the performance of the new global fitting was lower than that of PENTAtrainer even in the cross-validation condition, where each speaker's prosody is generated by parameters trained on all other 7 speakers.

Table 6: Mean RMSE & Pearson correlation reported in Xu [20] for all the speakers in the same English corpus

Learning mode \ Accuracy	RMSE	Correlation
Speaker dependent	2.07	0.836
Group average	2.77	0.772
Cross validation	2.98	0.757

The comparison of Table 6 and Table 2 is not entirely appropriate, however, because the full potential of AM theory is by no means adequately reflected by the current version of AMtrainer. Clearly missing are composite tones like LH* and HL*, phrase tones like H- and L-, baseline or reference line, and possibly pitch range specifications [9]. Also missing is parameterization of tone alignment [4，10], which is potentially critical for naturalness. For these nuanced aspects of the theory, however, we will need insights or direct collaborations from researchers with greater familiarity with AM theory than us.

5. CONCLUSIONS

This paper has presented the first version of AMtrainer that can perform predictive synthesis with categorically parameterized tone targets in American English. Despite the simplicity of the parameterization (global averaging of F_0 heights), fairly natural sounding F_0 configurations could be generated that bear both focus and sentence type information, showing that AM Theory has some promise as a model for predictive speech synthesis. This is a step forward toward full quantification of AM theory. It also further enriches the Common Prosody Platform [16].

These results also have interesting implications about AM Theory. Since the AM labels are abstract representations of tones, and abstractions must be made concrete in order to make predictive synthesis possible. To this day, AM tones are labeled manually based on observed F_0 events. So the theory provides no way to predict where exactly tonal targets will actually occur. To fully computationalize the AM Theory, theorists must at least agree on where the AM tones should occur based on theoretical predictions. More work is therefore needed in future research.

6. REFERENCES

[1] AMtrainer (v. 1.3.9). Accessible at www.homepages.ucl.ac.uk/~uclyyix/AMtrainer/

[2] Anderson, M., Pierrehumbert, J. and Liberman, M. 1984. Synthesis by rule of English intonation patterns. In *Proceedings of Proceedings of ICASSP*, San Diego, CA: 77-80.

[3] Arvaniti, A. (in press). The autosegmental metrical model of intonational phonology. In *Prosodic Theory and Practice*. J. Barnes and S. Shattuck-Hufnagel. Cambridge: MIT Press pp.

[4] Arvaniti, A. and Ladd, D. R. (2009). Greek wh-questions and the phonology of intonation — Discussion of major differences between AM and PENTA. *Phonology* 26 (01): 43-74.

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

[6] Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* **5:9/10**: 341-345.

[7] Hirose, K., Hashimoto, H., Saito, D. and Minematsu, N. (2016). Superpositional modeling of fundamental frequency contours for HMM-based speech synthesis. In *Proceedings of Speech Prosody 2016*, Boston, USA**:** 771-775.

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

[9]Ladd, D. R. (2008). *Intonational phonology*. Cambridge: Cambridge University Press.

[10] Ladd, D. R., Schepman, A., White, L., Quarmby, L. M. and Stackhouse, R. (2009). Structural and dialectal effects on pitch peak alignment in two varieties of British English. *Journal of Phonetics* **37**(2): 145-161.

[11] Lee, A., Xu, Y., Prom-on, S. 2014. Modeling Japanese F0 contours using the PENTAtrainers and AMtrainer. TAL 2014. Nijmegen**:** 164-167

[12] Liu, F., Xu, Y. 2007. Question intonation as affected by word stress and focus in English. In *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1189-1192). Saarbrücken: International Congress of Phonetic Sciences.

[13] Mixdorff, H. (2012). The application of the Fujisaki model in quantitative prosody research. In *Understanding Prosody – The Role of Context, Function, and Communication*. O. Niebuhr. New York: Walter de Gruyter pp. 55-74.

[14] Pierrehumbert, J. 1980. The phonology and phonetics of English intonation (Doctoral dissertation, Massachusetts Institute of Technology).

[15] Pierrehumbert, J. 1981. Synthesizing intonation. *Journal of the Acoustical Society of America* 70: 985-995.

[16] Prom-on, S., Xu, Y., Gu, W., Arvaniti, A., Nam, H., Whalen, D. H. 2016. The Common Prosody Platform (CPP) — where Theories of Prosody can be Directly Compared. In *Proceedings of Speech Prosody 2016*, Boston, USA**:** 1-5.

[17] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of The 1992 International Conference on Spoken Language Processing*, Banff**:** 867-870.

[18] Xu, Y. 2004. Transmitting tone and intonation simultaneously-the parallel encoding and target approximation (PENTA) model. In International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages.

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

[20] 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.