Comparative Modelling of Speech Prosody: AM Theory vs. PENTA Model

Albert Lee[1] & Faith Chiu[2]

[1]The Education University of Hong Kong (Hong Kong), [2]University of York (United Kingdom)
albertlee@eduhk.hk, faith.chiu@york.ac.uk

Several rival models of speech prosody have coexisted for several decades. In the case where these competing models contribute to mutually exclusive proposals, it becomes necessary for the prosody researcher to directly compare and assess the models and the theories behind each of them. Having originated by accounting for phenomena from different language families, each of these models was thus proposed to serve different purposes (e.g. with the main focus of representing theoretical grammatical constructs, or to generate f_o curves identical to what is observed acoustically). As a result, the models also tend to vary on their fundamental assumptions, their f_o generating mechanisms, and the number and levels of input specification required. These differences have made it difficult for researchers to fully understand rival models well enough to assess them fairly, thus further contributing to their continuous coexistence.

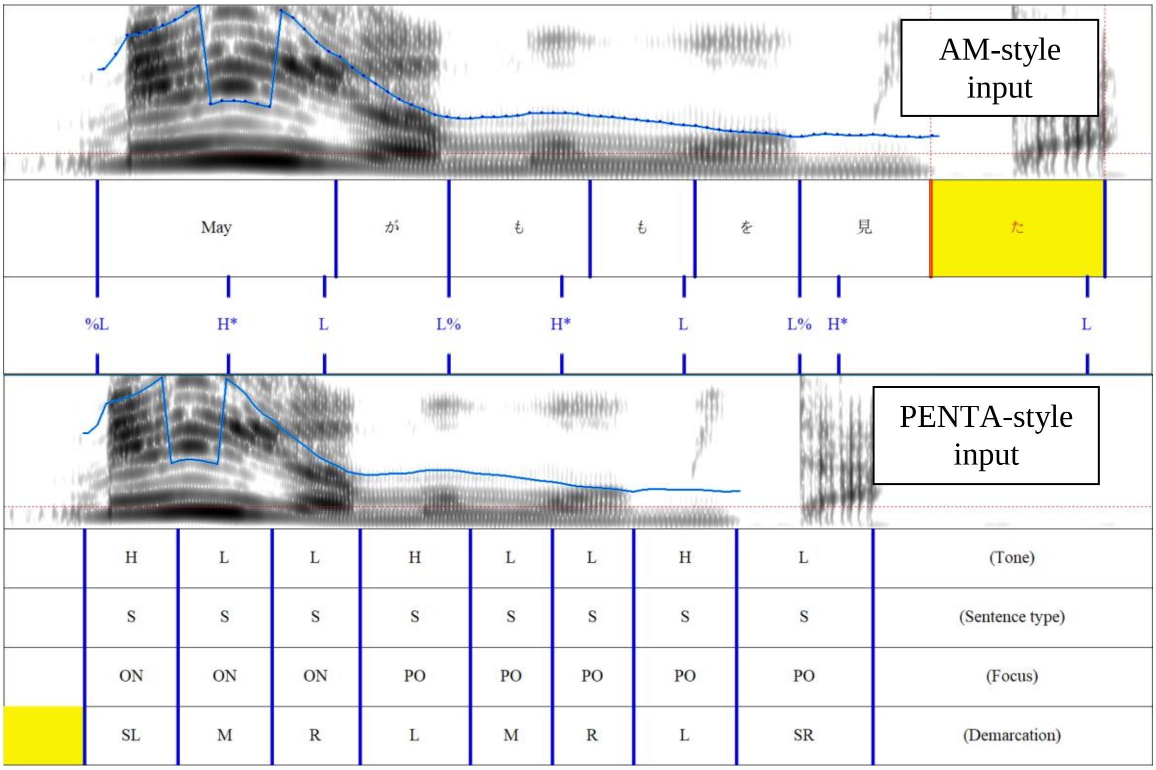
Comparative modelling is one objective way of directly comparing models. By resynthesizing a given dataset based on the workings of several theories, their respective synthesis accuracy can serve as a gauge for fair assessment. For example, the Common Prosody Platform [1] is a recent endeavor intended to offer a user-friendly platform for such comparison.

In this paper, we propose four factors that users should take into consideration when comparing theoretical models based on synthesis accuracy. They are (i) the underlying targets that generate surface f_o contours, (ii) the level of target specification, (iii) the degree of freedom permitted in the number of tiers which contributes to information encoding of each target, and (iv) how they implement underlying units into generation of f_o contours.

To illustrate these points, we present the synthesis accuracy performance of a 6,400-sentence corpus of Japanese utterances [2] that contrasts in narrow focus condition (initial / medial / final / neutral), sentence type (declarative / interrogative), and lexical accent condition (initial accent / unaccented). We implemented speaker-dependent resynthesis and speaker-independent predictive synthesis based on two models: the Autosegmental-Metrical Theory (AM henceforth, e.g. [3]) and Parallel Encoding and Target Approximation Model [4].

We generated AM-styled and PENTA-styled annotation files (see examples overleaf) using a PRAAT script. These files were manually checked and rectified before being submitted to speaker-dependent resynthesis. Subsequently, we used the Jackknife procedure [5] to carry out speaker-independent predictive synthesis. On the whole, it was found that PENTA yielded better synthesis accuracy than AM. We take this to argue that PENTA's (iv) f_o-generating mechanism can achieve a better curve fit.

Nevertheless, AM and PENTA also differ in terms of (i) the nature of underlying targets, (ii) level of target specification, and (iii) degree of freedom in the number of tiers permitted to encode information. As it stands, there is no fair way to quantitatively compare models in these three aspects. It is concluded that the Common Prosody Platform has taken a crucial step forward for prosody research, with much that remains to be achieved, and that there are aspects of theories that awaits further developments before they could be assessed by computational or mathematical means.



	H	L	L	H	L	L	H	L	(Tone)
	S	S	S	S	S	S	S	S	(Sentence type)
	ON	ON	ON	PO	PO	PO	PO	PO	(Focus)
	SL	M	R	L	M	R	L	SR	(Demarcation)

Fig.1 AM-style and PENTA-style annotation of speech data in preparation for model training.

References

[1] 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 the 8th International Conference on Speech Prosody (SP2016)* (pp. 1–5). Boston, MA.

[2] Lee, A., &Xu, Y. (2018). Conditional realisation of post-focus compression in Japanese. In *Proceedings of the 9th International Conference on Speech Prosody (SP2018)* (pp. 216–219). Poznań, Poland.

[3] Pierrehumbert, J. B., &Beckman, M. E. (1988). *Japanese Tone Structure*. Cambridge, MA: Massachusetts Institute of Technology.

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

[5] Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*, 353–360.