Predicting Taiwan Mandarin tone shapes from their duration

Chierh Cheng[1], *Michele Gubian*[2]

[1] Department of Speech, Hearing and Phonetic Sciences, University College London, UK
[2] Centre for Language and Speech Technology, Radboud University, Nijmegen, NL

Chierh.Cheng@googlemail.com, M.Gubian@let.ru.nl

Abstract

A preliminary study on modelling tonal variation as a function of duration is carried out. An experimentally controlled acoustic database was utilized to construct functional linear models. In the construction of the linear models, duration was used as independent variable in predicting the shape of disyllabic pitch contours in Taiwan Mandarin, given the target tone sequences. Results showed that by moving duration values from short to long, tonal curve shapes of disyllables ranging from non-reduced to reduced were approximated with an adequate goodness-of-fit (usually below one semitone RMSE). This study provides a novel approach to examine the relation between duration and F_0 realisation of small units such as disyllables and also supports the time pressure account of phonetic reduction in general.

Index Terms: tonal reduction, duration, functional linear models, Taiwan Mandarin, Functional Data Analysis (FDA)

1. Introduction

Work on connected speech has shown a close link between duration and the degree of phonetic reduction, i.e. short duration can coincide with high degrees of reduction and vice versa [6,10,13,18]. To investigate the continuous durational effect on phonetic reduction, this study will utilize an experimentally-controlled database of Mandarin tones contained in disyllabic nonsense words to construct functional linear models in which duration alone will be used to predict tone shapes, given the information of which tone combination is produced. This approach is motivated by Lindblom's model of durational undershooting, that is, the shorter the duration, the greater the target undershoot [9]. This study aims to test the hypothesis of an underlying gradual mechanism governing tonal reduction by providing evidence for the continuous nature of durational effects on phonetic reduction.

Taiwan Mandarin, the standard Chinese spoken in Taiwan, has four lexical tones: High (H,˥), Rising (R,˧˥), Low (L,˨˩) and Falling (F,˥˩). Research on the Chinese languages has highlighted a range of segmental and tonal changes that take place when a sequence of two or more syllables is concatenated under high time pressure [1,3,16]. For example, in Taiwan Mandarin, *wo <u>zhi dao</u>* [wo˨˩ tʂɨ˥ tau˥˩], 'I know' can be reduced into *wo <u>zhao</u>* [wo˨˩ tʂau] with a variant tone on the contracted syllables [tʂau]. In more extreme cases, trisyllables can also be reduced into monosyllabic units, such as *wo <u>bu zhi dao</u>* [wo˨˩ pu˥˩ tʂɨ˥ tau˥˩], 'I don't know' becoming *wo <u>bao</u>* [wo˨˩ pwau].

Previous research has shown evidence that extreme segmental reduction is the direct result of speaker's attempt to achieve the underlying segments under severe time pressure [4]. It has also been demonstrated that when two syllables are merged into one, speakers still appear to make an effort to produce the original tones within the contracted syllable [5]. In this study, we further test whether it is possible to predict the F_0 contour shape of a given sequence of two tones only from its duration. To test this, we first collect speech production data experimentally so as to eliminate the influence of discourse context. We then construct and evaluate a set of linear models that accept a single real number as predictor (duration) and output a whole contour, expressed in the form of a function of time. This kind of model is called Functional Linear Model (FLM), which is one of the techniques available in the family of advanced statistical techniques called Functional Data Analysis (FDA) [20,21]. Results show that despite their simplicity, these models can approximate F_0 contour shapes with an adequate goodness-of-fit, even in case of reduced tokens.

Below, we briefly describe the acoustic data collected and the methods used (i.e. FLM) and then present a series of modelling analyses and discuss the implications for the mechanisms of tonal reduction.

2. Methodology

2.1. Database

Six male native Taiwan Mandarin speakers aged 21-28 were recorded. The phonetic material used to construct the linear models consisted of disyllabic /ma/+/ma/ nonsense sequences with a total of 16 (4x4) tone dyads, which were embedded in two carrier sentences (16 tone dyads follow H and L tones respectively, where a dominant carry-over influence from the preceding tone was considered [22]), resulting in a total of 32 sets of tone combinations, i.e. H#HH, L#HH, H#HR, … , H#FF and L#FF. Various conditions were imposed to elicit different degrees of tonal reduction, i.e. position of the intermediate phrases occurred in the carrier sentence (initial, mediate and final), repetition time (1st, 2nd, and 3rd) and speech rate (slow, normal and fast). More details regarding the design of the material can be found in [5].

Reduction types were manually labelled according to the integrity of the intervocalic /m/, where items were labelled as *non-reduced* when an intervocalic nasal murmur was clearly present, *reduced* when the nasal murmur was clearly absent and *semi-reduced* in intermediate cases. Labelling decisions were taken by the first author (who is a native Taiwan Mandarin speaker) based purely on the presence/absence of intervocalic nasal murmur but not on duration. Among all speakers, two speakers (C and K) had both non-reduced and reduced items in all of their respective 32 sets of tone combinations (i.e. relatively balanced counts in each set). Thus, data from those speakers was selected to construct the linear models for a general analysis.

Extraction of F_0 contours was first carried out with the vocal cycle marking of the Praat program [2] and then with manual repair of octave jumps and other distinct irregularities using a Praat script [23]. For each target curve, 20 measurement points were generated, 10 equidistant points per syllable for *non-reduced* tokens (hence in total 20) and 20

equidistant points for *reduced* tokens. F_0 values are converted to semitones and the average of the 20 samples subtracted from all contours. This helps reduce variability owing to speaker identity and makes the estimation of functional linear modelling more straightforward.

2.2. Functional Data Analysis

2.2.1 Functional Linear Models

Functional Data Analysis (FDA) was introduced in the late 90's by J. Ramsay and colleagues [20,21]. All FDA techniques allow one to do statistical analysis on a set of contours (F_0 in our case) using only the information contained in their shape. Quantitative analysis of F_0 contours and other time-varying quantities (e.g. formants) are usually carried out by selecting a few shape features, such as peak and valley coordinates, slopes etc., and then using standard statistical analysis on the derived fixed-length feature vectors [14]. This approach forces one to choose in advance which shape features are relevant and which are not. Often feature extraction is carried out by hand. Another approach is to use a model such as the Fujisaki model [8] or the qTA model [17] for F_0. These are powerful models that take into account the physiology of phonation. Their performance depends on how faithfully the larynx or the vocal tract is modeled and how well the parameter tuning is carried out—and of course models for F_0 cannot be directly applied to formants or other signals. In contrast, FDA is a flexible platform that allows one to (i) use sampled contour values directly as input to the statistical analysis and (ii) refrain from introducing hypotheses on the nature of the analysed signal. The output of FDA is based solely on the regularities found within the set of input contour shapes.

Functional Linear Models (FLM) extends ordinary linear models to accept functions (of time) as input and/or output. The FLM version applied here takes a real number as independent variable (predictor) and predicts a whole contour shape, expressed as a function of time. In our case, the predictor d is a convenient transformation of the duration D of a disyllabic unit, while the output $f(t)$ is the predicted F_0 contour shape defined on a *fixed* time interval. In this way the shape and duration d are decoupled. Formally we have:

$$f(t) = \beta_0(t) + \beta_1(t) \cdot d, \qquad (1)$$

where d is the logarithm of the normalised duration D, that is $d = \log(D / \bar{D})$ where $\bar{D}$ is the average across all measured durations of tokens in the model. $\beta_0(t)$ and $\beta_1(t)$ are the functional parameters to be estimated, the analogous of the scalar parameters estimated in ordinary linear regression. The training of model (1) takes place similarly to the way ordinary linear regression models are trained (see [20] for details). The main difference for the user is that each training element is a $(d, f(t))$ pair, while the original F_0 contours are sampled. Hence the sampled F_0 contours have to be represented by continuous functions of time $f(t)$ before training can take place. Moreover, they have to be modified to cover the same time interval because the functions $f(t)$, $\beta_0(t)$ and $\beta_1(t)$ must be defined on a common interval [11].

2.2.2 Data Preparation

The problem of choosing a function $f(t)$ that best fits a set of samples is solved by applying standard smoothing techniques [20,21]. The user must choose a basis function, which for non-periodic signals is typically a B-spline. The internal parameters of the B-spline basis and the degree of smoothing imposed on the curve fitting were empirically determined by generalized cross-validation [20,21]. Examples of the quality of the smoothing process can be seen in Figure 3 and Figure 4: Dots indicate the original F_0 samples and the dashed lines their respective functional representation. To obtain an (apparent) constant duration for each curve, a fictitious [0, 1] normalised time interval is simply divided into 20 constant intervals, i.e. one per sample point. In this way, the functions $f(t)$ will be scaled so that the half curves spanning each syllable (in the *non-reduced* case) will be aligned with the center of the interval. This improves the analysis quality in that it takes away variability due to random misalignment of syllables. All FDA operations were carried out using the freely available R package "fda" [19].

2.2.3 Evaluation

In order to test the hypothesis of an underlying gradual mechanism governing tonal reduction, a model such as (1) was built for each tone combination. The rationale is that once the tone sequence is known, model (1) can be a simple yet adequate description of a gradual shape adjustment rule controlled by the available amount of time for production. In practice, models had to be further specialised by building separate models for different preceding tones (L or H) and also for the two selected speakers (C and K), since the simple structure of (1) cannot accommodate for the influence of those factors. Regarding speaker dependency, we recall that model (1) is trained on the surface realisation of a number of F_0 contours, while no physiological parameter estimation takes place. In total, 16 (tone dyads) x 2 (preceding tones) x 2 (speakers) = 64 FLM were produced. Each model was trained on around 25 (log duration, smoothed F_0 curve) pairs fairly balanced in the proportion of *reduced*, *semi-reduced*, *non-reduced* samples.

The 64 models were evaluated in terms of goodness-of-fit on their training set. For each model we computed the root mean squared error (RMSE) and the R^2 coefficient of determination averaged on their training set. R^2 is defined as:

$$R^2 = 1 - SSE/SSY, \qquad (2)$$

where SSE is the sum of squared errors resulting from approximating the sampled F_0 values with $f(t)$ in (1), SSY the sum of squared errors from fitting a horizontal line at a height corresponding to the average F_0 value, i.e. fitting a horizontal line is taken as a baseline goodness-of-fit [15]. Note that R^2 is not the square of anything and it takes a negative value when the predicted curve $f(t)$ makes a larger squared error than just fitting a horizontal line (i.e. when SSE > SSY).

3. Results and Discussion

Figure 1 and Figure 2 display the R^2 and RMSE value distributions across the 64 models in form of boxplots. The leftmost columns show R^2 and RMSE values measured when predicting all the contours belonging to a specific tone-speaker combination. The two middle columns show the same, but separately for non-reduced and for reduced contours respectively (results for semi-reduced contours are not shown separately but they are included in the first column). At least half of the models exhibit acceptable goodness-of-fit, which is remarkable if we consider the simplicity of (1) and the small number of training contours per model. The small differences between goodness-of-fit for reduced and non-reduced contours can be explained by the fact that reduced curves are flatter than non-reduced ones, thus the gain of fitting (1) relative to fitting a horizontal line (i.e. the baseline error considered in R^2) tends to be smaller for reduced contours, hence a smaller

R^2 value in the reduced case. On the other hand, RMSE tends to be smaller (i.e. better) in the reduced case because reduced curves have smaller range, so errors tend to have smaller absolute values.

At this point we wanted to test whether it is in fact easy to predict the shape of a reduced tone combination because tones may simply be flattened and all tone combinations become similar to each other. To verify this we have tried to predict reduced contours using mismatched models. The rightmost columns in Figure 1 and Figure 2 display goodness-of-fit when predicting reduced contours using a model randomly picked from among the 31 models trained on the same speaker but on another tone combination/context. The large absolute and relative performance deterioration shows that reduced F_0 contours still preserve information in their shape and they are not just flat. This confirms that the performance of (1) in predicting reduced contours is *not* due to an F_0 flattening occurring uniformly across the board. In particular, comparing the 3rd and the 4th columns in Figure 1, we see that fitting a matched model (1) to predict a reduced contour is better than a flat line ($R^2 > 0$) in 75% of the cases (i.e. from the lowest line of the box to the top), while the opposite is true when fitting a mismatched model (1).

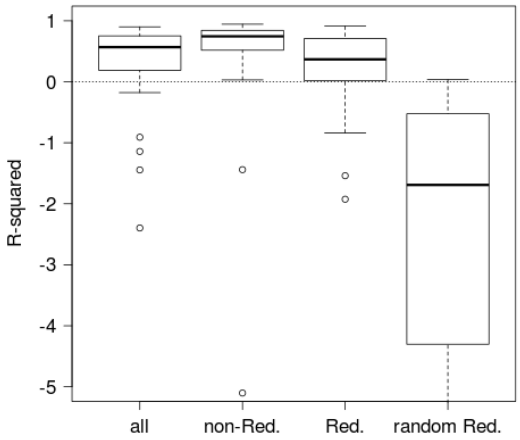

Figure 1: *R^2 values of comparing observation and prediction, from left to right, all contours, non-reduced contours, reduced contours and reduced contours with mismatched models.*

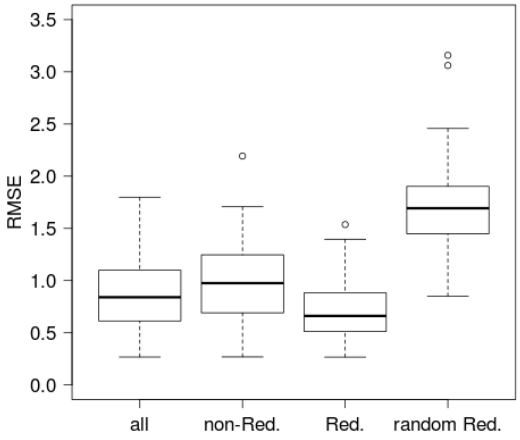

Figure 2: *RMSE values of comparing observation and prediction (from left to right, same order as in Figure 1).*

After providing global results, here we show some specific examples in the attempt to gain more insight. Figure 3 and Figure 4 show some contours selected from the model constructed for speaker K in the L#RH set, which globally scored RMSE = 0.35 and R^2 = 0.80. Figure 3 shows cases of good fitting whereas Figure 4 shows cases of poor fitting in this model. In each plot, the x-axis is the normalised time from 0 to 1 (so it is not in seconds) and the y-axis measures semitones (note the mean F_0 has been removed from each curve). Non-reduced curves are in thin black, reduced curves in thick orange. The twenty points of measurement for each curve are shown as dots and the dashed lines are the smoothed contours produced for training the model. The predicted curves are represented as solid lines. In Figure 3, the cases of good fitting, we see an adequate approximation when the predictor of duration changes from longer (D = 0.402s, d = 0.109, black) to shorter (D = 0.212s, d = -0.169, orange). This picture suggests evidence for the continuous nature of durational effects on phonetic realisation, since the reduced F_0 contour seems to preserve some traits of the non-reduced one, and this in turn is nicely captured by the FLM (1).

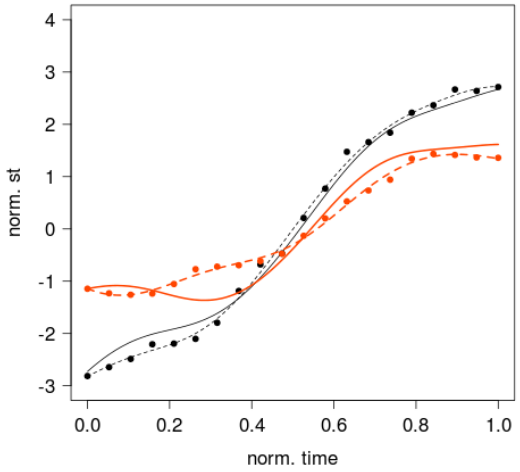

Figure 3: *Cases of good fitting of the model constructed for Subject K and tone set as L#RH. Point of measurements are shown as dots, smoothed f(t) shown as dashed curves. Solid curves are the respective predictions: Thin black for non-reduced and thick orange for reduced one.*

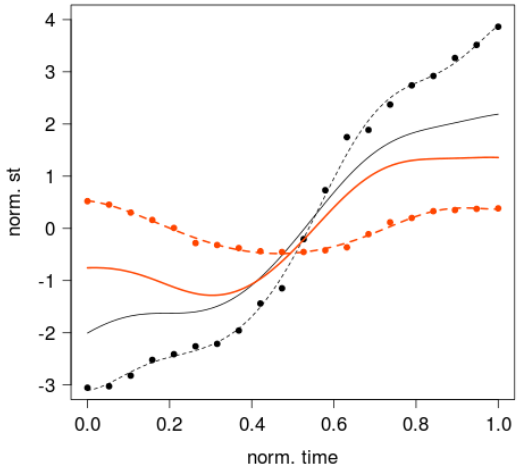

Figure 4: *Cases of poor fitting (same condition as in Figure 3)*

In Figure 4, the cases of poor fitting (using the same model as in Figure 3), we see that the chosen non-reduced curve (black dashed) is of greater amplitude than its prediction (D = 0.300s, d = -0.017, black solid) but still of similar shape, i.e. rising from low to high. On the other hand, the predicted reduced curve (orange solid) shows a general mismatch in terms of shape and range. We attribute this failure to the limited predicting power of (1), which uses only one predictor, *d*. When time pressure is as high as in our cases of contracted syllables (D = 0.131s in the example), speakers simply cannot realise each target in time (though still attempted to), thus producing a seemingly 'flattened' contour [5]. By visually inspecting several other cases we found patterns similar to those exemplified above.

Combining this and the reported goodness-of-fit results we conclude that there seems to be evidence for the continuous nature of durational effect on tonal reduction, because a model as simple as (1) can capture a substantial part of it. However, when reduction is too extreme, i.e. speakers may reach their physiological limit, models in predicting phonetic variation shall take into account speakers' physiological mechanisms.

4. Conclusions

An experimental and data-driven approach was adopted to investigate the close link between duration and tonal reduction. With the support of functional linear modelling applied to an experimentally-controlled acoustic database, we found evidence for the continuous nature of durational effects on phonetic reduction, which is consistent with Lidnblom's duration-dependent target undershoot model [9]. It is concluded that the mechanism behind tonal reduction is governed by an articulatory process (i.e. a continuous mechanism [7]) and the speakers' inherent physiological limit needs to be considered when modelling extreme reduction as shown in the examples presented here [8,12,17,24].

5. Acknowledgements

The research of Michele Gubian was supported by the Marie Curie Research Training Network Sound2Sense (www.sound2sense.eu). More on Michele Gubian's research on FDA can be found at his website: lands.let.ru.nl/FDA. We would like to thank Yi Xu and Lou Boves for reading the draft and making a number of helpful suggestions, although of course we are solely responsible for any errors.

6. References

[1] Berry, J., "Tone reduction in Mandarin Chinese", Journal of Acoustical Society of America, 125(4): 2571-2571, 2009.

[2] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer [Computer program]", Version 5.2.21, retrieved from http://www.praat.org/, 29 March, 2011.

[3] Cheng, C. E., "An acoustic phonetic analysis of tone contraction in Taiwan Mandarin", MA, National Cheng Chi University, Taipei, 2004.

[4] Cheng, C. and Xu, Y., "Extreme reductions: Contraction of disyllables into monosyllables in Taiwan Mandarin", In Proceedings of Interspeech 2010., Brighton, 456-459, 2009.

[5] Cheng, C., Xu, Y. and Gubian, M., "Exploring the mechanism of tonal contraction in Taiwan Mandarin", In Proceedings of Interspeech 2010, Makuhari, 2010-2013, 2010.

[6] Flege, J. E., "Effects of speaking rate on tongue position and velocity of movement in vowel production", Journal of the Acoustical Society of America, 84(3): 901-916, 1988.

[7] Farnetani, E. and Recasens, D., "Coarticulation and connected speech processes", in W. J. Hardcastle, J. Laver and F. E. Gibbon [Eds], The Handbook of Phonetic Sciences, 2nd Edition, 316-352, Wiley-Blackwell, 2010.

[8] Fujisaki, H., Wang, C., Ohno, S. and Gu, W., "Analysis and synthesis of fundamental frequency contours of standard Chinese using the command-response model", Speech Communication, 47: 59-70, 2005.

[9] Lindblom, B., "Spectrographic study of vowel reduction", Journal of the Acoustical Society of America, 35: 773-1781, 1963.

[10] Lindblom, B., "A note on segment duration in Swedish polysyllables", Speech Transmission Laboratory Quarterly Progress Status Report, 2: 1-5, 1964.

[11] Gubian, M., Boves, L. and Cangemi, F., "Joint analysis of F0 and speech rate with Functional Data Analysis", To appear in the Proceeding of International Conference of Acoustics, Speech and Signal Processing 2011, ICASSP, 2011.

[12] Kochanski, G. and Shih, C., "Prosody modeling with soft templates", Speech Communication, 39: 311-352, 2003.

[13] Moon, S. and Lindblom, B., "Interaction between duration, context, and speaking style in English stressed vowels", Journal of the Acoustical Society of America, 96: 40-55, 1994.

[14] Morén, B. and E, Zsiga., "The lexical and post-lexical phonology of Thai tones," Language, 24(1): 113-178, 2011.

[15] Motulsky, H and Christopoulos, A., "Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting", GraphPad Software Inc., San Diego CA, www.graphpad.com, 2003.

[16] Myers, J. and Li, Y. S., "Lexical frequency effects in Taiwan Southern Min syllable contraction", Journal of Phonetics, 37: 212-230, 2009.

[17] Prom-on, S., Xu, Y. and Thipakorn, B., "Modelling tone and intonation in Mandarin and English as a process of target approximation", Journal of the Acoustical Society of America, 125: 405-424, 2009.

[18] Vatikiotis-Bateson, E., and Kelso, J. A. S., "Rhythm type and articulatory dynamics in English, French and Japanese", Journal of Phonetics, 21: 231-265, 1993.

[19] R Development Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org, 2011.

[20] Ramsay, J. O. and Silverman, B. W., "Functional data analysis (2nd Edition)", Springer, 2006.

[21] Ramsay, J. O., Hookers, G. and Graves, S., "Functional data analysis with R and MATLAB", Springer, 2009.

[22] Xu, Y. "Contextual tonal variations in Mandarin", Journal of Phonetics, 25: 61-83, 1997.

[23] Xu, Y., "_ProsodyPro.praat", online from 2005-2011: http://www.phon.ucl.ac.uk/home/yi/tools.html, accessed on 29 Mar, 2011.

[24] Xu, Y. and Prom-on, S., "Articulatory-functional modeling of speech prosody: A review. In Proceedings of Interspeech 2010, Makuhari, 2010.