Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 51 (2018) 44–67

www.elsevier.com/locate/csl

A hierarchical linguistic information-based model of English prosody: L2 data analysis and implications for computer-assisted language learning☆

Chao-yu Su[a,b,c,*], Chiu-yu Tseng[b], Jyh-Shing Roger Jang[d], Tanya Visceglia[e]

[a] *Institute of Information Systems and Applications (ISA), National Tsing Hua University, Taiwan*
[b] *Institute of Linguistics, Academia Sinica, Taiwan*
[c] *Taiwan International Graduate Program, Academia Sinica, Taiwan*
[d] *Department of Computer Science & Information Engineering, Taiwan University, Taiwan*
[e] *Freelance writer and independent researcher*

Received 30 November 2017; received in revised form 29 January 2018; accepted 3 March 2018
Available online 14 March 2018

Abstract

The paper presents a prosody model of native English (L1) continuous speech as corrective prosodic feedback for non-native learners. The model incorporates both hierarchical discourse association and information structure to (1) pinpoint the prosodic features of multi-phrase continuous speech, and (2) simulate native-like expressive speech using corpus of North American and Taiwan L2 English. The bottom-up, additive, data-driven model aims to generate L1-like expressive continuous speech with built-in phonetic and phonological specifications at the lexical level, syntactic/semantic specifications at the next higher phrase and sentence levels, and completed with patterned paragraph associations and prosodic projections of information allocation at higher levels. The hierarchical model successfully allows us to identify L1-L2 differences by prosodic modules/patterns as novel additional features "discourse structure" and "information density" reliably nail down L1-L2 prosodic differences related to phrase association as well as information placement. Our L1 prosodic model with the proposed predictors and optimized model trained from L1 speech corpus showed increase of prediction over existing methods. As a corrective feedback for L2 learners, these predicted L1 prosodic features were compared with a baseline model by objective evaluation (RMS error and correlation) then superimposed onto the L2 speech tokens. Resynthesized L2 tokens were subsequently compared with the original L2 tokens for degrees of perceived accent using subjective evaluation (native-listener perception test). We believe the proposed model can be an effective alternative for implementing computer-assisted language learning (CALL) systems that helps generate L1-like prosody from text, and at the same time serves as corrective feedback for L2 learners.

Keywords: Prosody; Discourse structure; Information structure; L2 english; Resynthesis; CALL; linguistic; Continuous speech

☆ This paper has been recommended for acceptance Roger K. Moore.

[*] Corresponding author at: Institute of Information Systems and Applications (ISA), National Tsing Hua University, Taiwan.
E-mail address: morison@gate.sinica.edu.tw (C.-y. Su).

http://dx.doi.org/10.1016/j.csl.2018.03.001
0885-2308/ 2018 Elsevier Ltd. All rights reserved.

1. Introduction

The goal of the present study is develop a prosody model that integrates discourse association and information structure to simulate continuous L1 English speech prosody towards more comprehensible L2 communication. The model is intended as CALL baseline for advanced L2 learners who need to produce more intelligible and expressive continuous speech by improving overall global prosody. Our motivation stems from that fact that though quite a number of individual prosodic features have been identified to contribute to L2 accent, intelligibility and comprehensibility, little is known as to how these features are interactively related to more expressive speech, and where L2 speakers' learning attention could be directed. However, we noted that most of existing CALL systems tended to use prosodic features derived from data of read isolated sentences instead of more realistic continuous speech, and inadvertently by design left issues specific to continuous speech prosody unaddressed. Our group has long been investigating speech prosody with data of continuous speech, and adopted a top-down hierarchical perspective from the start. We consider this departure of perspective from the mainstream is particularly significant when dealing with fluent continuous speech consisting large size multi-phrase speech unit whose output generation involves both linear and hierarchical derivation, and by default would bring to light prosodic issues of global nature. The following presentation reports our current attempt to CALL application that was encouragingly supported by recent works (Domínguez et al., 2014, 2016) that not only accounted for the close relationship between discourse prosody and information structure, but also predicted discourse prosody in continuous speech. We believe works devoted to analyzing and simulating more realistic sentences of continuous speech, whether well planned or spontaneous, merit more research attention.

In the physical sense, prosody is generally referred to the melodic and rhythmic aspects of speech that involves modulations of fundamental frequency, duration, and amplitude in the speech signal (Munro, 1995; Scruton, 1996; Derwing et al., 1997; Anderson-Hsieh et al., 1992; Benrabah, 1997; Coniam, 1999; Witt et al., 2000; Trofimovic et al., 2006; Moustroufas et al., 2007). By examining the acoustic aspects of speech output at face value it is no surprise that some reported studies concluded that the range of prosodic variation is non-systematic and unpredictable (James et al., 1976; Peppé et al., 2000; Jacewicz et al., 2010). However, we believe this is largely due to examining output acoustic data at face value instead of understanding its composition that involves both linear association as well as hierarchical governing. Note that three parallel layers contribute to prosody formation, namely, linguistic, para-linguistic, and nonlinguistic. The linguistic layer encodes phonetic representation, lexical (semantic), syntactic (phrase and sentence), discourse information and some pragmatic information that can be predicted from content. The paralinguistic layer conveys speaker's attitudes, emotions, dialect, sociolect, idiolect, etc. The non-linguistic layer delivers the speaker's gender, age and physical state. In other words, the majority of output prosodic variations from linguistic contributions should be largely predictable, and therefore could be modeled. For instance, the procedures of how the syllables are combined to form prosodic words and how prosodic words form prosodic phrases and sentences are already well known (Selkirk, 1984; Nespor and Vogel, 1986). Furthermore, various prosodic hierarchy supported by other L1 studies also verified how acoustic correlates at each prosodic level contribute collectively to surface prosody, as well as identifying a range of communicative functions such as marking stress, focus, and boundaries etc. (Bailly et al., 2005; Fujisaki et al., 2005; Xu, 2005; Mixdorff, 2002a). The fact is acoustic models that successfully separate contributions from words and sentences (Laver, 1991; Fujisaki, 2004) are available; these models could easily be extended to accommodate additional higher level contributions from larger size speech units. For instance, instead of examining single spoken phrases or simple sentences at one time, Tseng et al. (2005, 2008), as an alternative, extended a perceived prosodic hierarchy to include discourse-level multi-phrase association patterns in order to account for the formation and generation of cross-phrase global prosody of continuous speech from multiple levels of contributions. Through corpus studies of Mandarin L1 speech data, the perception based discourse hierarchy, supra-segmental acoustic correlates F0, duration, intensity and pause duration were analyzed with a regression procedure that teased apart each prosodic correlate from surface prosody into particular discourse levels as well as their cumulative contribution to output prosody. It turned out that the hierarchical alternative not only accounted for contributions from the syllable, prosodic word, and prosodic phrase, but also successfully took into account physiological constraint change of breath when speaking continuously and multi-phrase units, thereby substantiating contributions above words and sentences in order to form coherent multi-phrase speech paragraphs. Thus a working model capable of predicting the interaction and modulation among involved discourse layers and ultimate prosody output prosody of multiple-phrase units is already in existence to account for how underlying prosodic patterns systematically correspond to various discourse levels and derivationally contribute to output prosody.

In comparison to the relatively little corpus linguistic investigation of output prosody that accounts for phrase/discourse association, even less attention has been paid to prosodic modulation conditioned by arrangement of information structure beyond the sentence level. While information structure (IS) coded in words (lexical) as well as larger structure (sentence) generally refers to the organization that reflects the important content of utterances (Halliday, 1967; Chafe 1974; Lambrecht, 1994), utterance in these studies remained at the sentence level and corpus based results scarce. Among the two best accepted definitions of IS from L1 studies, the more generally assumed one is the given/new dichotomy of information status (Prince, 1981) while the less assumed other is focus structure (König, 2002). We note here that as working definition for corpus analysis, the sentence-based definition of given/new status soon became inadequate when used to analyze data of continuous speech, for example, narratives because the same nouns and noun phrases would appear over and over. Graphic display of acoustic data of continuous speech also showed how the intonation contours of individual phrases varied highly on the one hand, and almost always contained more than one peak on the other. Instead, reported works showed how prosodic features in English, Bulgarian, Italian and Dutch are attributed to focus types (Hoskins, 1997; Sityaev and House, 2003; Avesani and Vayra, 2003; Hanssen et al., 2008; Andreeva et al., 2016), hence focus specified status appeared to be a more possible and plausible working definition to analyze information structure of continuous speech data. We therefore tested perception of focus types as a separate level of manual annotation for data preprocessing, and were able to retain consistent tagging across transcribers. Consequently, we adopted focus type as our working definition of Information Structure.

We noted that similar to L1 studies, the majority of L2 studies have also been limited to single specification by isolated single tokens at a time, at individual levels and without mention of cross-level and cross sentence interactions. For example, at the phonetic level, studies of L2 English consonants and vowels produced by Javanese and Swedish speakers showed how their temporal patterns differ from L1 English due to influences from their respective mother tongue (Thorén, 2007; Perwitasari et al., 2015). At the lexical level, when L2 English of Japanese and Mandarin speakers was compared with L1 English, it was found the L2 speakers produced weaker acoustic contrasts between stressed and unstressed syllables (Nakamura 2010; Tseng et al., 2013). But there has been little to no report on the phonetic-lexical interaction in perceived L2 accent, let alone discourse features that must be taken into account to produce continuous speech of multiple phrase units.

Similar to reported works on L1 prosodic features, information structure related L2 prosodic features has also received less attention, and attention has also been paid to sentence and discourse prosody separately. At the sentence level, most reported results showed that expressing information structure via prosody turned out to be more challenging than expected for L2 speakers. For example, given/new information related pitch accent placement of L2 English speech by German, Spanish, Japanese, Malay and Thai speakers showed that L2 speakers would often emphasize given information instead (Wennerstrom, 1994; Grosser, 1997; Ramirez Verdugo, 2002; Gut, 2009, 2013). Focus structure related pitch accent placement, i.e., broad vs. narrow focus, by L1 Taiwan Mandarin and Beijing Putonghua speakers showed insufficient differentiation of on-focus/post-focus contrasts (Visceglia et al., 2012). Vietnamese and Hong Kong L2 English exhibited similar weakened realization of pitch accent contrast (Nguyễn et al., 2008) and post-focus compression (Gananathan et al., 2015). L2 focus realization distinct from L1 was also found in Atterer and Ladd (2004) and O'Brien and Gut (2010) at the sentence level. At the discourse level, patterns of chunking and phrase association that may attribute to L2 comprehensibility/accent have also been studied, though not as extensively as sentence level issues. Both L1 Taiwan Mandarin (Tseng et al., 2010) and Bengali (Saha and Mandal, 2017) speakers exhibited similar chunking features: (1) Inconsistency of realizing discourse-level chunking, continuation or termination among speech paragraphs and among speakers. (2) More units of intermediate chunking than L1 English speakers. In short, these studies collectively demonstrated that though some distinct features related to individual prosodic units and/or levels have been widely studied, the interaction among them was hardly addressed, especially with respect to discourse/paragraph association and information structure. The lack of understanding of how individual features interact is also evidenced by a recent comprehensive review of CALL systems applying TTS (Text To Speech) that concluded by recommending existing CALL systems to pay more attention to the development of natural prosody and expressiveness (Handley, 2009).

Motivated by the common lack of interactive studies to help determine the prosodic constitution of multi-phrase speech units with appropriate information placement, especially Computer-Aided Language Learning (CALL) systems for more advanced L2 learners, we set our goal to construct a prosody training system for CALL applications that would incorporate interaction of involved factors and trained with data of continuous speech. We hope the linguistic information-based model of English prosody could bring more implications for advanced computer-assisted language learning. In the following sections, we will present our proposed model based on models summarized in Bailly et al. (2005), Fujisaki et al. (2005), Xu (2005), and Tseng et al. (2005) using refined methods from earlier research (Zellner et al., 2001; Tseng et al., 2005, 2008). We will use a hierarchically inclusive perspective integrating linguistic-layer categories from both discourse and information structure to analyze L2 prosody. Specifically, prosodic modules/patterns were extracted from surface prosody at particular linguistic levels in order to provide a finer-grained, hierarchical analysis of the individual and collective contributions made by each prosodic level, as well as to investigate the possibility of interaction among those levels. Based on data-driven approaches, a bottom-up, additive model of L1 prosody was built, starting with phonetic and phonological specifications at the lexical level; then superimposing higher-level syntactic/semantic specifications at the phrase and sentence levels. After which patterned prosodic projections of paragraph associations and information structure were added to produce fluent continuous speech. The same model was also used to compare L1 and L2 Taiwan English prosody from a hierarchical perspective, which allowed us to identify differences in production of prosodic modules/patterns at each level of linguistic specification, as well as the interactions among these levels. We also developed a L1 prosodic model to provide corrective norm for L2 learners by simulating L1 prosodic features using the proposed predictors and optimized model trained from L1 speech corpus. Simulated L1 prosodic features were compared with a baseline model by objective evaluation (RMS error and correlation). The simulated L1 prosodic features were further superimposed onto L2 speech tokens, resynthesized and compared with original L2 tokens in terms of perceived accented using subjective evaluation (native-listener perception test). We will show that the increased prediction accuracy and reduced L2 accent makes the model a good candidate for CALL implementation, in particular how it could be used as corrective feedback toward prosody training.

2. Corpus and annotation

2.1. Corpus design

Spoken-language tasks were excerpted from the AESOP_ILAS database, whose materials were designed to elicit a range of supra-segmental features, based on previous research investigating the prosodic contribution to perception of accent in L2 speech (Visceglia et al., 2009). These tasks included 20 frequency-controlled and stress-balanced (2−4 syllable) target words (Appendix A) in the following prosodic contexts: (1) in 20 carrier sentences (Appendix B) (2) at prosodic boundaries (Appendix C) (3) in the position of contrastive stress (Appendix D). The design attempts to elicit prosodic patterns of (1) canonical lexical stress in the target words as well as following superimposing/higher-level prosodic alteration of (2) prosodic boundaries from the syllable, word, phrase to sentence level and (3) focus status at sentence level in the context where 20 target words are embedded. Speakers were also required to produce the passage ‘The North Wind and the Sun’, which allowed measurement discourse-level prosodic features/linguistic specifications in longer units of speech. Recorded speech from 11 L1 speakers of North American English (5 M/6F) and 30 speakers of Taiwan L2 English (15 M/15F) were used. The Taiwan L2 speakers were asked to self-rate their English proficiency by 4 levels, namely, poor, average, good and excellent. The distribution of reported proficiency level is listed in Table 1. Total of 660 spoken sentences/11 speech paragraphs of L1 English and 1800 spoken sentences/30 speech paragraphs in Taiwan L2 English were selected.

2.2. Annotation for linguistic specifications

Five types of linguistic specifications related to prosody were annotated for the purpose of analysis and modelling, namely, segmental information, lexical stress pattern, syntactic structure, information structure (focus type) and

Table 1
Distribution of proficiency level by self-evaluation of Taiwan L2 speakers.

Proficiency level	Poor	Average	Good	Excellent
Percentage	20.00	50.00	26.67	3.33

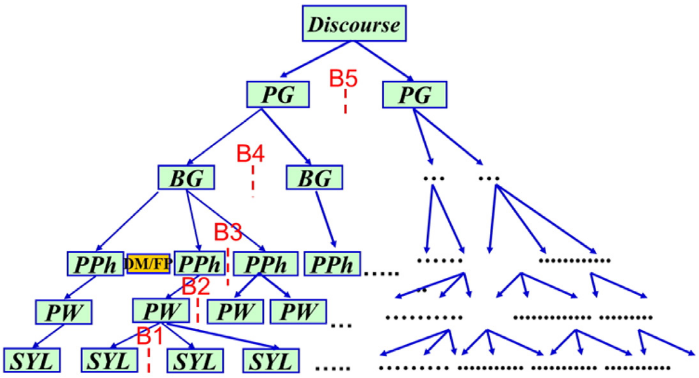

Fig. 1. A hierarchical diagram of perceived discourse structure.

discourse structure. Segmental identities were automatically tagged on an audio timeline using the HTK Toolkit followed by manual spot-checking by trained transcribers; perceived boundaries were manually tagged while transcriber consistency over 80%. Three levels of word/lexical stress primary (P), secondary (S) and tertiary (T) were automatically tagged in syllable units using the CMU electronic dictionary. Syntactic structure was annotated at syllable, word and phrase boundaries by a native American English linguist. Phrase boundaries were further categorized into non-phrase boundary (NB), continuation rise (CR), final rise (FR) and final fall (FF). Information structure by narrow focus (NF), broad focus (BF), non-focus (NonF) and function word (FW) was annotated by the same native American English linguist. Examples of the above annotation on AESOP_ILAS are provided in Appendix B, C and D.

Hierarchical discourse structure was perceptually annotated on an audio timeline without consideration of textual marks such as punctuation. The hierarchical discourse structure was annotated into five levels of prosodic units: the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG, a physio-linguistic unit corresponding to change of breath while speaking continuously) and the multi-phrase speech paragraph (PG). These units were manually tagged as 5 levels of perceived discourse boundary B1 through B5 (Tseng et al., 2005). The unit/boundary correlations can be expressed as SYL/B1, PW/B2, PPh/B3, BG/B4 and PG/B5 as shown in Fig. 1. A between-transcriber consistency rate for discourse boundaries in the training corpus of 80% or above was required for the transcriber to annotate the present corpus.

The following presents an example of a complete perceived PG with text in ‘The North Wind and the Sun’ produced by a male native English speaker.

$|_{\mathbf{B5}}$**Then the sun shone out warmly**$|_{\mathbf{B4}}$**, and immediately the traveler took off his cloak**$|_{\mathbf{B4}}$**. And so the north wind**$|_{\mathbf{B3}}$ **was obliged to confess that the sun was the stronger of the two**$|_{\mathbf{B5}}$**.**

In the example, although ‘**warmly**’, ‘**his cloak**’ are respectively completions of phrase and sentence in the text, the respective following prosodic boundaries are perceived as clear continuations with a breath change (B4) while the containing speech paragraph lasts until ‘**of the two**’ (B5). These perceived boundaries are not simply silence, but audio cues to listeners of how phrases are associated to form semantic coherence/cohesion. In turn, each and every phrase within the phrase group is not the same when they are produced individually and in isolation. An annotation example combining all levels of annotations by discourse structure and information structure is illustrated in Fig. 2.

3. Feature extraction

Three acoustic features of surface prosody are broken down into particular linguistic specifications for prosodic analysis and prediction. The acoustic features include magnitude of accent/phrase command (Aa/Ap) for pitch analysis and phoneme duration (PD) for tempo analysis. After Aas/Aps are extracted, Aas and Aps are respectively aligned into perceived SYL and PPh units (Annotation) to analyze and predict the Ap and Ap modules in relation to each level of linguistic specifications which were also aligned respectively into SYL and PPh units. Extraction of acoustic features and definition of linguistic specifications are presented in the present section. More details for division of prosodic modules into linguistic levels of specification are given in Section 4.1. L1 prosody was then

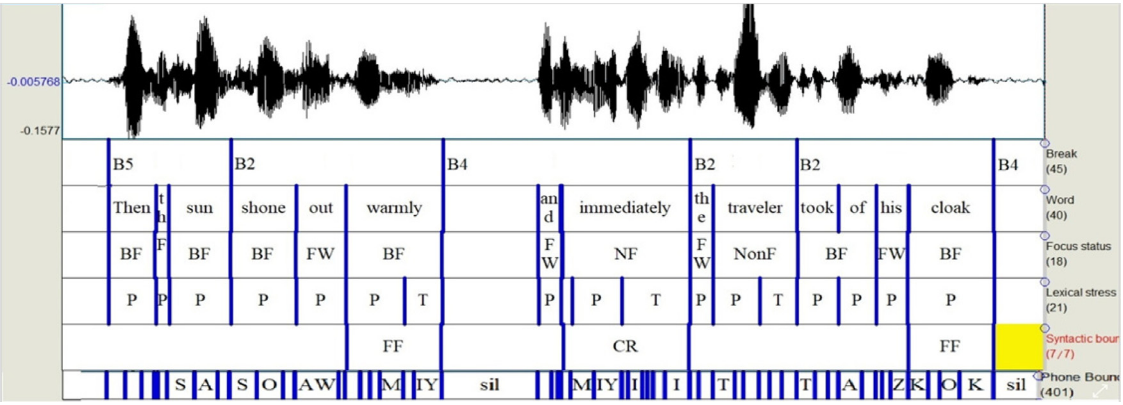

Fig. 2. An annotation example encoding both discourse structure and information structure in “the north win and the sun”.

reconstructed by optimizing interaction among each level of prosodic modules using regression models, which will be discussed in Section 4.2

3.1. Acoustic/prosodic variables

3.1.1. Accent and phrase commands Aa/Ap (F0)

Accent/phrase commands Aa/Ap were extracted using the command-response model (Hirose et al., 1984) which decomposes the surface f0 contour into three components: the speaker’s base register (base frequency, $\mathrm{F_b}$), long-term/global tendency (phrase component, $\mathrm{A_{P*}G_P(t)}$) and short-term/local humps (accent component, $\mathrm{A_{a*}G_a(t)}$). The three components and corresponding commands and parameters can be seen in the sequence (1)−(3) below. Analysis and modeling in the present study focused on the two parameters that dominate F0 high\low contrast, namely the accent magnitude (Aa) and phrase command (Ap). Our method for automatic extraction of Ap and Aa is presented in the following section.

$$F = \ln(F_b) + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\left[G_a\left(t - T_{1j}\right) - G_a\left(t - T_{2j}\right)\right] \tag{1}$$

$$G_p(t) = \alpha^2 t \exp(-\alpha t),\ for\ t \geq 0 \tag{2}$$

$$\begin{aligned} G_a(t) = {} & \min[1 - (1+\beta t)\exp(-\beta t),\ \gamma], \\ & for\ \ t \geq 0\ ,\ \ \text{where}\ \ \alpha = 3,\ \ \beta = 20 \end{aligned} \tag{3}$$

3.1.1.2. Auto-extraction of Aa/Ap. The most commonly used auto-extraction of parameters in Eqs. (1)−(3) for tone languages are based on low-pass filtering (Mixdorff, 2000); this method is better suited for tone languages because it derives both positive and negative Aas. In order to extract only positive Aas for a non-tone language, such as English, we developed a simpler extraction method that still uses local minima of low-pass contours (Mixdorff, 2000) for intonation contour segmentation. After intonation boundaries were segmented, optimized Ap values were determined using grid search which aims to find minimal distance between phrase component in Eq. (2) and original F0 contour at some significant F0 points including first local peak and the following several local minimum of original F0 within the current intonation segment. These significant F0 points, in which local minima are prevalent, result in an optimized contour of phrase contour passing through most of local minima of original F0 contour. The optimized phrase contour in low values means residual components obtained from original F0 subtracting the optimized phrase contour would be positive. The positive residual components which function as approximation targets of Aa function in (3) produced positive Aa values. A demonstration of auto-extraction for Ap is shown in Fig. 3.

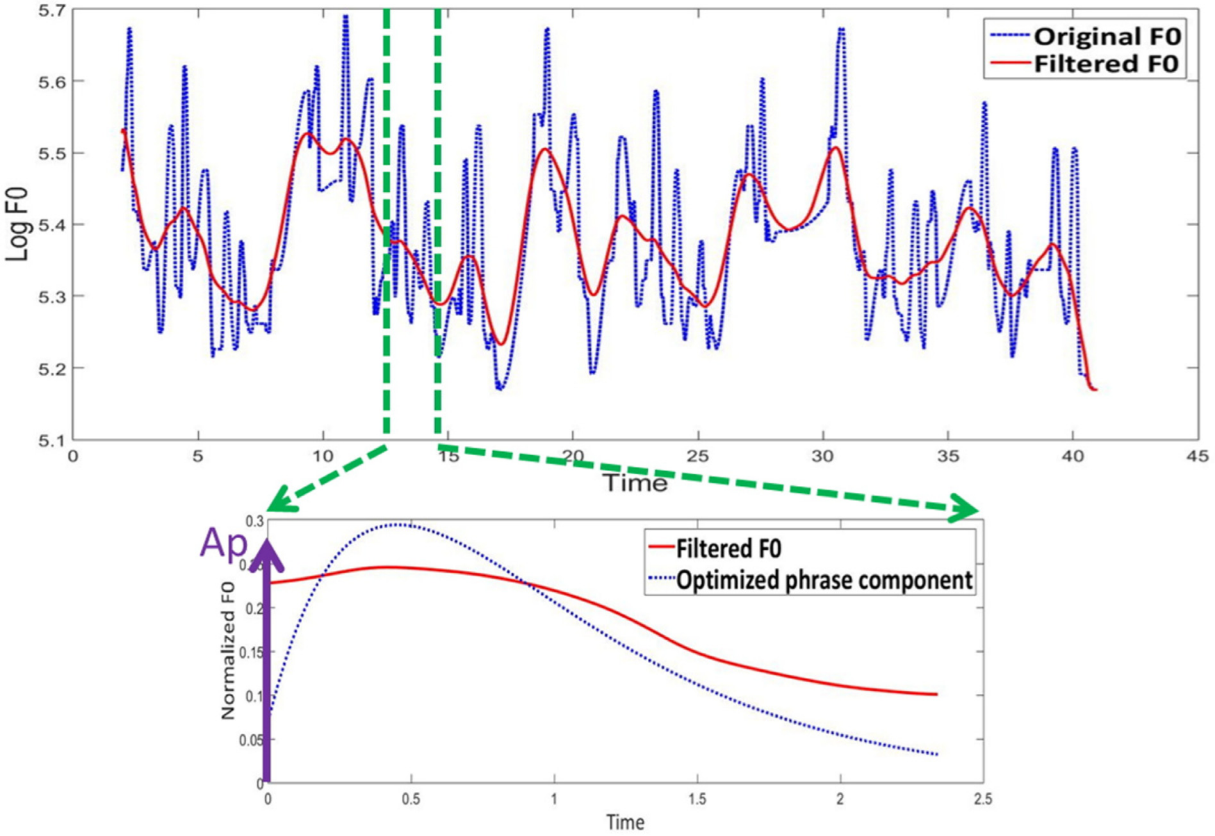

Fig. 3. An example of auto-extracted phrase component (lower) from entire F0 contour of 'The North Wind and the Sun' (upper). Most of extracted Aps correspond to PPh units described in Section 2.2.

After Aas/Aps extracted, Aas and Aps are respectively aligned into perceived SYL and PPh units (Section 2.2) to analyze and predict the Ap and Ap modules in relation to each level of linguistic specifications which were also respectively aligned into SYL and PPh units (will be shown in Section 3.2).

3.1.2. Phoneme duration (PD)

PD was directly extracted at phoneme boundaries which are force-aligned by the HTK Toolkit followed by manual spot-checking (Section 2.2). Extracted Aa, Ap and PD were then aligned with the linguistic specifications outlined in the following section for further analysis and modeling.

3.2. Linguistic variables

The linguistic specifications "phonological" (stress), "information planning" (focus structure), "syntactic" (sentence modality) and "phrasing" (segmentation) were extracted by annotation (see Section 2.2 for details) and distributed across Ap, Aa and PD for analysis and modeling. Contextual specifications were extended based on the differences in scale and other properties of Ap, Aa and PD outlined in the sections below.

3.2.1. Linguistic variables for modeling Aa

Based on previous studies suggesting prosody is a layering outcome of multiple linguistic levels based on Bailly et al. (2005), Fujisaki et al. (2005), Xu (2005), Mixdorff (2002a), and Tseng et al. (2005, 2008), Table 2 lists possible linguistic predictors for modeling Aa at different levels. Among the predictors, type of lexical stress (X1), focus degree (X5) and boundary type (X10) are basic variables at the lexicon, sentence and discourse level.

Table 2
Summary of linguistic variables for Aa modeling.

Affecting level	Feature	Code
Lexicon level	Lexical stress type of current syllable	X1
	Pre-syllable contrast within a word	X2
	Post-syllable contrast within a word	X3
	Relative syllable position by primary stress in a word	X4
Sentence level	Focus degree of current word	X5
	Pre-word contrast within a sentence	X6
	Post-word contrast within a sentence	X7
	Relative position to NF in a sentence	X8
	Relative position to first BF in a sentence	X9
Discourse level	Boundary type (level) of current phrase	X10
	Boundary type (level) of pre phrase	X11
	Boundary type (level) of post phrase	X12
Information structure	Infodensity by syllable	X13
	Infodensity by word	X14

Starting from the lexicon level, an initial Aa modules are directly derived from annotations of lexical stress (X1). For instance, all Aa values corresponding to primary stress in target words (Appendices A and Appendices B) are grouped and averaged to represent initial Aa models of primary stress. However, such extraction is inappropriate for higher-level linguistic specifications. For instance, Aa values derived for focus degree at sentence level (X5) using the same extraction intrinsically comprise lower-level effects such as lexicon stress (X1) and higher-level effects such as focus degree (X5). To clearly separate individual level contributions of X1 and X5 before modeling, an additional separation refined from (Zellner et al., 2001; Tseng et al., 2005; Tseng et al., 2008) is proposed and explained in detail in section Section 4.1.

After the Aa modules are separated into each disjointed prosodic level/linguistic specification (Section 4), the contextual features extended from X1, X5 and X10 (X2-X4, X6-X9, X11-X12) are derived in a contrastive perspective. For instance, pre-and post- syllable contrast (X2 and X3) is defined as extracted Aa values in the current syllable position subtracting those in pre- or post- syllable as demonstrated by Eqs. (4) and (5).

$$AX2_i = AX1_i - AX1_{i-1} \tag{4}$$

$$AX3_i = AX1_i - AX1_{i+1} \tag{5}$$

where i is the position index of current syllable. $AX1_i$, $AX1_{i-1}$, $AX1_{i+1}$ *respectively represent the extracted Aa values in current, preceding and following syllable position.*

Relative positions to information centers including (1) the primary-stress syllable in a word and (2) a narrow-focused word in a sentence are also considered as factors for Aa prediction (Tseng et al., 2014). Tseng et al. showed that position relative to information centers is a predictor for systematic/optimal patterns of L1 prosody whereas L2 features are less sensitive to the same factors. The relative positions to information center are defined as demonstrated in Eq. (6).

$$\mathrm{X7} = i - \ i_{NF} \tag{6}$$

where i and i_{NF} *are respectively the absolute position indexes of current word and narrow focused word.*

A simple illustration of relative positions to second type of information center, namely narrow-focused word in a sentence, is demonstrated as follow.

YOU SHOULD ***TAKE***$_{(BF)}$ *THE*$_{(NonF)}$ ***ELEVATOR***$_{(NF)}$ *INSTEAD OF THE STAIRS.*

In the example, the position of ‘THE’ relative to the information center is defined as the absolute position of ‘THE’ in the current sentence subtracting the absolute position of narrow focus, namely ‘ELEVATOR’, in the current sentence. Because ‘THE’ and ‘ELEVATOR’ are respectively the 4th and 5th words in the current sentence, the obtained position of ‘THE’ relative to the information center is −1, 4−5 = −1.

Another important linguistic level, i.e., information structure (X13/X14), is derived from the labels of focus status smoothed using a defined function, information density, as shown in Eq. (7). Information density determines the information weight at the current position by averaging the weights of nearby individual small units such as words/syllables. The function is based on an assumption that larger-scale planning for neighborhood information context is more physiologically efficient than smaller-scale planning for individual units, which may lead to frequent changes of articulator.

$$X14_i = \frac{1}{2n+1} \sum_{i=-n}^{n} X5_i \tag{7}$$

where i is the current word position index and n represents scale setup for surrounding context (n = 1 setup in the following example, n = 2 fine tuned in results Section 5.1.4).

A simple illustration of information density is provided as follows:

*YOU SHOULD **TAKE**$_{(BF)}$ THE$_{(NonF)}$ **ELEVATOR**$_{(NF)}$ INSTEAD OF THE STAIRS.*

In the example, the information weight of ‘THE’ is jointly decided by ‘TAKE’, ‘THE’ and ‘ELEVATOR’ . If non-focus, broad focus and narrow focus information weights are respectively setup as 1, 2, 3, the information weight of ‘THE’ would be interpolated from 1 to 2, (2 + 1 + 3)/3.

3.2.2. Linguistic variables for modeling Ap

Higher-level discourse and information planning specifications for Ap modeling are listed in Table 3. Similar to extracting Aa modules described in Section 3.2.1, Ap values are grouped and averaged by the variables in Table 3 to derive Ap modules into each level of linguistic specifications. Again, module separation between levels before Ap modeling is also required to make clear the interactive effect from lower to higher levels listed in Table 3. For Ap modules at discourse level such as PG, PG positions should be normalized due to various PG lengths by PPh/Ap number. The normalized PG positions are defined in *Eq. (8)*.

$$NorP_{PG} = i_{Ap}/N_{PG} \tag{8}$$

where $NorP_{PG}$, i_{Ap} and N_{PG} respectively represent normalized PG position, absolute position index of current Ap in current PG and total number of Ap in current PG.

The normalized PG positions are further quantized into 6 bins for grouping Aps and deriving the Ap average.

After the PG effect is subtracted from the surface Ap, the residual Aps are also grouped and averaged by normalized information density to derive the additive Ap modules of information density. The normalized information density is defined by Eq. (9).

$$NorID_{PPh} = N_{BF+NF}/L_{PPh} \tag{9}$$

where $NorID_{PPh}$, N_{BF+NF} and L_{PPh} respectively represent the normalized information density of a phrase, the number of BF plus NF in current phrase and the total word number in current phrase.

The normalized information density is quantized into 5 bins for grouping Aps and deriving the Ap average.

3.2.3. Linguistic variables for modeling phoneme duration (PD)

A total of four linguistic variables were used as input features for PD modeling: phoneme identity, type of lexical stress, focus degree and syntactic structure. Phoneme identity was included as a linguistic variable for modeling PD, in order to feature intrinsic durational difference among phoneme types.

Table 3
Summary of linguistic variables for Ap modeling.

Discourse structure	Phrase length by word number Break level preceding a phrase Break level following a phrase Normalized phrase position in a BG Normalized phrase position in a PG Distance to pre phrase by word number
Information allocation	Number of BF and NF per phrase Number of NF per phrase NF position in a phrase Information density in a phrase

4. Methodology

4.1. Separation of acoustic/prosodic variables into levels of linguistic specification

Aa, Ap and PD were divided into explicit modules/patterns corresponding to linguistic levels, respective in order to determine the contribution made by each level to surface prosody. An Aa example given in (10) shows the layered contribution from each level of linguistic specifications listed in Table 2 to surface Aa.

$$A_{SF} = M_{X1} + M_{X2} + M_{X3} + \ldots \ + M_{X14} \tag{10}$$

where A_{SF} represents surface Aa and $M_{X1} - M_{X14}$ respectively represent the prosodic modules of Aa by each level of linguistic specifications listed in Table 2.

Derivation of particular prosodic module from surface Aa, taking stress labels (primary, secondary and tertiary) as example, is shown in (11) and (12). Note that (11) and (12) only use Aas extracted from the target words in carrier sentences (Appendix A and B) which is designed to elicit canonical prosodic patterns of lexical stress with minimal prosodic interaction from the other prosodic layers.

$$A_{SF} = \begin{bmatrix} A_1 \\ A_2 \\ \ldots \\ A\ _i \end{bmatrix}, \quad L_{X1} = \begin{bmatrix} L_1 \\ L_2 \\ \ldots \\ L\ _i \end{bmatrix}, \quad L_i \in \{\mathrm{ST}_j{:}'\mathrm{P}','\mathrm{S}',\ '\mathrm{T}'\} \tag{11}$$

where i, L_{X1} and j represent the observation index, corresponding target label (stress for example) and category index of the stress label;'P','S','T' respectively represent primary, secondary and tertiary stress.

$$M_{X1} = \begin{bmatrix} M_1 \\ M_2 \\ \ldots \\ M\ _i \end{bmatrix}, \quad M_i = \frac{1}{N_j} \sum A_{\{i:L_i \in \ ST_j\}} \tag{12}$$

where N_j is the number of the observations belonging to stress type ST_j.

Extracted Aa modules in the current level (which is stress, M_{x1} in example 13 below) were separated from surface prosody A_{SF} to derive higher level effect, A_H which represents the collective contribution from X2-X14.

$$A_H = \ A_{SF} - M_{x1} = M_{x2} + M_{x3} + \ldots \ + M_{x14} \tag{13}$$

If the focus status (X5) is chosen as the next target factor to separate, Aa modules at the other levels except for X5 (X2-X4, X6-X14) could be further teased apart by subtracting Aa modules at X5 from A_H . The derivation of Aa modules at X5 is the same as (11) and (12) while L_i replaced from stress labels (X1) to focus labels (X5). Note that, if focus status is targeted; only Aas extracted from tasks of contrastive stress (Appendix D) are used.

Applying the procedure of module extraction and exclusion from X1 to X14 iteratively, each level of prosodic modules could be separated from the surface prosody A_{SF} and analyzed.

Note that for each separation, only one individual target specification at a time is allowed in order to minimize possible interaction effects between specifications. This dynamic programming rationale can achieve final global optimization for each level of prosodic modules/patterns with minimal interaction from higher-level linguistic specifications.

4.2. Regeneration model of L1 prosody using linguistic specifications

After the Ap/Aa/PD modules had been separated into disjointed level, the whole Ap/Aa/PD set combining each level of modules could be served as a multi-level code book representing the prosodic hierarchy. According to the hierarchical code book of L1 prosody, textual annotations on scripts were transformed into corresponding prosodic modules at different level. Using the prosodic modules at each level as inputs, joint contribution from each level, which simulates mutual interaction among levels was optimized/trained by regression models. We assume the configuration is capable of regenerating the aggregate and interactive contribution of prosodic modules into native expressive prosody, which can provide corrective prosodic norm for the learner. The three types of regression models used are: Multivariable Linear Regression (MLR), Robust Regression (RoFit) and Feedforward Neural Network (FNN). These regression models simulate different types of mutual interaction among prosodic levels using linear combination (MLR/RoFit) or multi-layer nonlinear transformation (FNN). Training the regression models for optimizing joint contribution from each level is illustrated in Fig. 4. More details of the three types of regression model will be shown in Section 4.2.2.

According to observation for Aa deployment in real speech flow, Aas only appear in certain syllable positions, not in all syllable positions. Aas positions by syllable must be determined before predicted Aa values inserted. Whether current syllable position is inserted with Aa value is determined by decision tree (see Section 4.2.1) using predictors in Table 2 as inputs.

4.2.1. *Binary classifier for predicting Aa position*

To predict the absence/presence of Aa by syllable within a prosodic phrase, the linguistic variables in Table 2 were used as explanatory input to a decision tree: a predictive model with tree structures, which maps an item's explanatory inputs to a conclusion about that item's response values (Pedhazur, 1982). The split criterion used in this study is Gini's diversity index (gdi), with 'leave-one-out' added for model validation.

4.2.2. *Regression models for predicting Ap, Aa and duration*

Three types of regression models were used to predict three prosodic variables (Aa, Ap and PD) from the linguistic specifications in Section 3.2, including: Multivariable Linear Regression (MLR), Robust Regression (RoFit) and Feedforward Neural Network (FNN). MLR approximates the relationship between a response variable and a linear combination of explanatory variables (Utgoff, 1989). RoFit is an extension of multivariable linear regression, whose derived model is less sensitive to outliers (Andersen, 2008). FNN is a modeling technique for approximating a response variable using non-linear functions (Auer et al., 2008). The model used in this study includes thirty hidden layers.

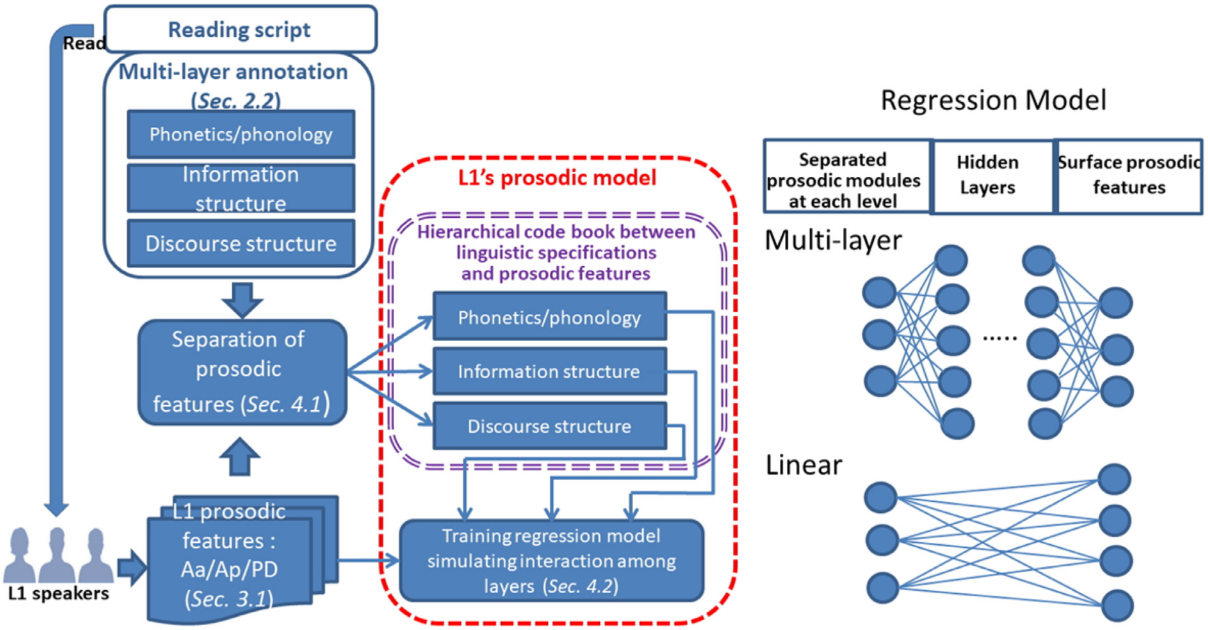
Read
Reading script
Multi-layer annotation (Sec. 2.2)
Phonetics/phonology
Information structure
Discourse structure
Separation of prosodic features (Sec. 4.1)
L1 prosodic features : Aa/Ap/PD (Sec. 3.1)
L1 speakers
L1's prosodic model
Hierarchical code book between linguistic specifications and prosodic features
Phonetics/phonology
Information structure
Discourse structure
Training regression model simulating interaction among layers (Sec. 4.2)
Regression Model
Separated prosodic modules at each level
Hidden Layers
Surface prosodic features
Multi-layer
.....
Linear

Fig. 4. Generating hierarchical code book of L1 prosody and training optimal model to simulate surface prosody of L1.

4.3. Resynthesis − generating corrective feedback for learners

Using the hierarchical code book of L1 prosody derived and the layering contribution optimized (Section 4.2.2), annotations on learning script for L2 learners could be transformed and resembled into simulated L1's patterns of F0 and tempo (Aa/Ap/PD). The simulated L1 patterns including continuous F0 contour recovered by transformed Aa and Ap, using Eq. (2) and (3), as well as duration features in phoneme units were than superimposed onto identical L2 sentences/paragraphs and resynthesized as corrective prosodic feedback for L2 learners. The resynthesis/transformation of L2 speech is implemented using the Time-Domain Pitch Synchronous Over Lap-and-Add (TD-PSOLA) (Malah, 1979) provided by Praat (http://www.fon.hum.uva.nl/praat/). An illustration is shown in Fig. 5. We assume that the resythesized L2 speech is comprised of basic naturalness due to the following two reasons. (1) In comparison to conventional speech synthesis that concatenates small units such as words into phrases and/or sentences and faces the issue of lack of naturalness head on, we choose to use L2 produced continuous speech of both short and longer complex sentences. As a result of design, our L2 speech is already more natural than word based synthesis output in the first place. Our attempt was to superimpose specific prosodic features to improve continuous speech to prosody only. (2) The current study sets up particular thresholds for each prosodic parameter to limit the degree of manipulation in order to rule out unnatural resynthesized speech due to extreme parameters. However, we believe parameter adjustment in relation to naturalness for such CALL feedback applications is an interesting issue that merits further study in the future.

4.3.1. Evaluation

4.3.1.2. Objective evaluation. The proposed model will be objectively evaluated by Root Mean Square Error (RMSE) and Pearson's correlation (Corr) between simulated and real/original L1 prosodic features. RMSE measures the difference between predicted and extracted prosodic features (Ap/Aa for F0 and PD for duration), and Corr measures the degree of their linear relationship. Furthermore, the RMSE and Corr in the current model are compared with a baseline model in a previous study (Mixdorff, 2002b). In the baseline model, several higher-level variables including prominence and sentence effect were considered and used as inputs of feedforward neural network for

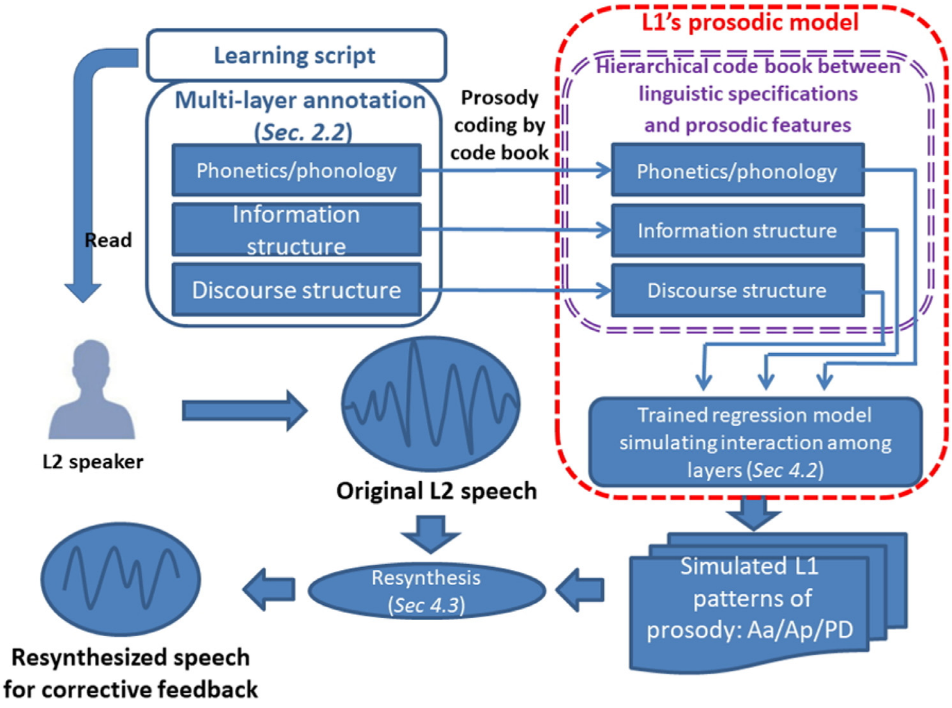
L1's prosodic model
Hierarchical code book between linguistic specifications and prosodic features
Phonetics/phonology
Information structure
Discourse structure
Trained regression model simulating interaction among layers (Sec 4.2)
Learning script
Multi-layer annotation (Sec. 2.2)
Phonetics/phonology
Information structure
Discourse structure
Prosody coding by code book
Read
L2 speaker
Original L2 speech
Resynthesis (Sec 4.3)
Simulated L1 patterns of prosody: Aa/Ap/PD
Resynthesized speech for corrective feedback

Fig. 5. Superimposing simulated L1 prosody onto L2's speech as corrective feedback for L2 learners.

Ap/Aa/duration prediction. Using the baseline model, high performance on Ap/Aa/duration prediction was achieved and reported.

The present study assumes that the proposed current prosodic model could improve baseline model in terms of prediction accuracy by three innovative features: (1) prosodic modules are separated into levels of linguistic specification, which are posited to be explanatory variables, whereas the baseline model have directly derived modules in a single flat level without separating mutual interaction between prosodic layers/linguistic specifications. (2) Information density was added to the predictor variables for Aa and Ap modeling. (3) Discourse-level specifications were added to the list of predictor variables for Ap. The relative improvement to baseline model will be reported in Section 5.3.1.

4.3.1.3. Subjective evaluation. Resynthesized L2 speech was subjectively evaluated by eight native English participants (4 M/4F) in a perception test to determine whether superimposition of L1 prosodic contours generated by the current prosodic model would make L2 speech more 'native-like' across a range of linguistic levels. Subjects were asked to answer three respective questions in Table 4. corresponding to three levels of improvement. Q: "Which target word/sentence/paragraph is more native-like?" A: (1) speech sample A, (2) speech sample B and (3) no difference. The examples A and B are randomly assigned into original and resynthesized L2 speech samples of the same text content. The goal is to subjectively evaluate whether resynthesized L2 speech using the present model is perceived as less accented than original L2 speech. The experiment included a total of fifteen perception subtasks, including twelve subtasks related to stress and focus, and three subtasks related to discourse.

5. Results

5.1. L1-L2 differences of F0 at some major linguistic specifications

5.1.1. Lexical stress

The L1/L2 F0 difference due to (lexical) stress status primary, secondary and tertiary is measured by degree of contrast whereby contrast is defined as subtraction of Aa values between the maximum and minimum, shown in Fig. 6 (Su et al., 2016a). Materials used to elicit these comparisons were designed to minimize the effects of other linguistic specifications, such as boundary effect, by embedding the target words in carrier sentences. Overall, Taiwan L2 English exhibits a lower degree of contrast among primary, secondary and tertiary stresses than L1 English, as shown in Fig. 6.

The L1 Aa patterns derived was recorded in the code book of L1 F0 at lexicon level to serve as lexicon-level inputs for prediction of L1 F0 (Section 4.2). The following prosodic modules/patterns derived will also be recorded to build up the hierarchical code book of L1 prosody from lower to higher level.

Table 4
Question set for perception test by linguistic level.

Lexical level	Which target word is more native-like?
Sentence level	Which sentence is more native-like?
Discourse level	Which paragraph is more native-like?

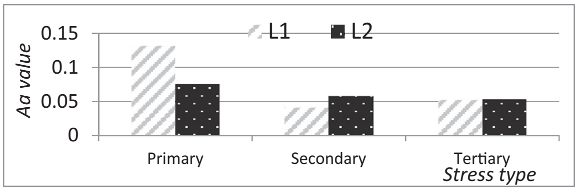

Fig. 6. Lexical-level Aa by stress type and speaker group; P, S and T = primary, secondary and tertiary stress.

5.1.2. Focus status

Fig. 7 shows Aa comparison by speaker group (L1/L2) and focus status after the effect of lexical stress (Fig. 6) was removed. Again, results indicate that Taiwan L2 English speakers produce weaker contrasts at the focus level than L1 speakers. Analysis of L1 patterns showed that relative larger additive Aa in narrow focus potion than broad focus and non-focus. The Taiwan L2 English speakers, in contrast, showed less differentiation of primary and tertiary stresses, and secondary stress departed even further from the L1 norms.

5.1.3. Discourse prosody

Fig. 8 shows patterns of Ap across PG positions in L1 and L2 speech (Su et al., 2016b). The L1 pattern shows global F0 declination across PG positions clearly, whereas the irregular L2 pattern exhibits no clear direction. L2 results suggest lack of discourse planning across PG positions.

5.1.4. Information allocation

Fig. 9 illustrates the relationship between Ap and information density within a PPh for both speaker groups. The L1 pattern shows an overall ascending tendency, indicating that information density increases across PPhs. The L2 pattern, in contrast, exhibits very weak correlation between Ap and information density.

5.1.5. Summary of F0 findings

The F0 study showed that by each specification L2 exhibited weaker contrast than L1. At the lower level, L2 speakers' produced lesser local F0 humps than L1; the lack of pronounced contrasts, therefore, can be regarded as an account of L2 accent. At the higher levels with respect to global F0 tendency in relation to discourse position and information density, we found a positive correlation between global F0 tendency and predictive variables, which

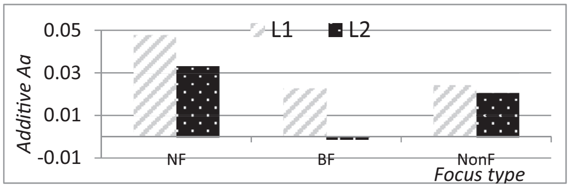

Fig. 7. Aa comparison by L1/L2 and focus status after the effect of lexical stress was removed; NF, BF and Non-F = narrow, broad and non- focus.

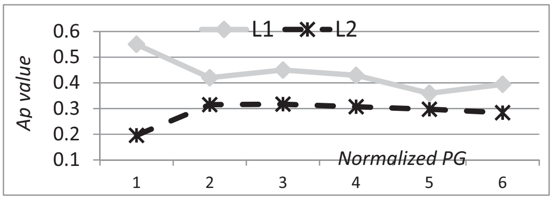

Fig. 8. Ap across PG positions in L1 and L2 speech. Vertical axis = Ap; horizontal axis = normalized and quantized PG position.

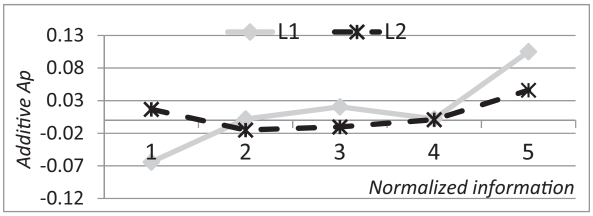

Fig. 9. Distribution of Ap by information density in L1 and L2 speech. Vertical axis = Ap; horizontal axis = normalized and quantized information density.

was absent in L2 speech. For PG position, L1 speakers produced global paragraph F0 declination, a prosodic feature which forms semantic association and cohesion, from the beginning of a PG to its end, whereas the inverse F0 pattern was found in L2 data. Hence, the F0 results from both lower and higher levels allowed us to compare the L1–L2 difference in a hierarchical/additive/finer-grained perspective which better accounts for accented L2 prosody as opposed to previous L2 studies that investigated individual levels without considering their interaction.

5.2. L1–L2 differences of tempo from linguistic specifications and boundary effects

5.2.1. Intrinsic segmental duration

The greatest L1/L2 difference in vowel duration was found in ʌ, ʊ, o, ɑ, and ə, for which between-group differences, ranging from 0.352 to 0.571 in normalized duration scale (Su et al., 2016c). As for consonants, the greatest L1–L2 difference was found in *θ*, ʒ, dʒ, h, and ŋ, ranging from 0.419 to 0.789 in normalized duration scale.

5.2.2. Lexical stress

Fig. 10 shows duration patterns by speaker group (L1/L2) and stress type after segmental effects was removed (Su et al., 2016c). L2 speakers produced a lower degree of contrast among primary, secondary and tertiary stresses than L1 speakers.

5.2.3. Boundary cues

Fig. 11 shows duration patterns by boundary type and speaker group (L1/L2) after lower-level effects from segmental duration and lexical stress was removed. With the exception of NB, all L1 patterns show considerable preboundary lengthening; the degree of lengthening was almost identical among CR, FR and FF (0.173, 0.172 and 0.171). In Taiwan L2 English, however, a considerably smaller contrast degree of lengthening was found across all boundary types, particularly in type CR (L1: 0.173, L2: −0.024).

5.2.4. Focus marking

Fig. 12 compares L1 and L2 production of the duration patterns of marking focus status after removal of segmental, lexical and boundary effects (Su et al., 2016c). Focus categories include narrow focus (NF), broad focus (BF), non-focus (NonF) and function word (FW, subcategorized from NonF). The focus status of syllables was further divided into primary, secondary and tertiary stress types, as the adjustments in segment duration used to mark stress may vary according to stress types (see Fig. 10). In the L1 primary and secondary stress patterns demonstrated (Fig. 12), lengthening increased with focus status in the order of NonF < BF < NF, whereas tertiary stress was

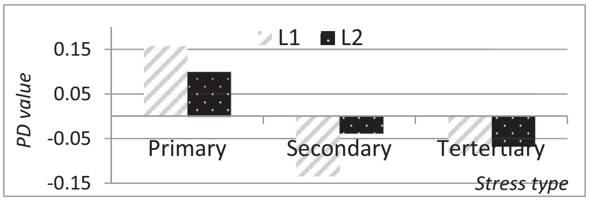

Fig. 10. Duration patterns by stress type and speaker group (L1/L2) after subtraction of segmental effects.

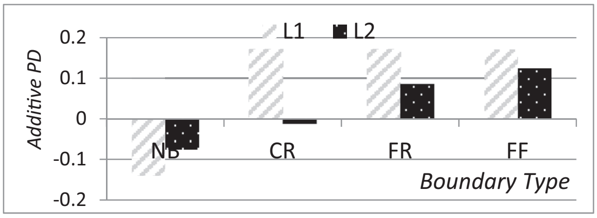

Fig. 11. Duration patterns by boundary type and speaker group (L1/L2) after subtraction of segmental and stress effects. Non-phrase final boundary, continuation rise, final rise and final fall are labeled NB, CR, FR, and FF, respectively.

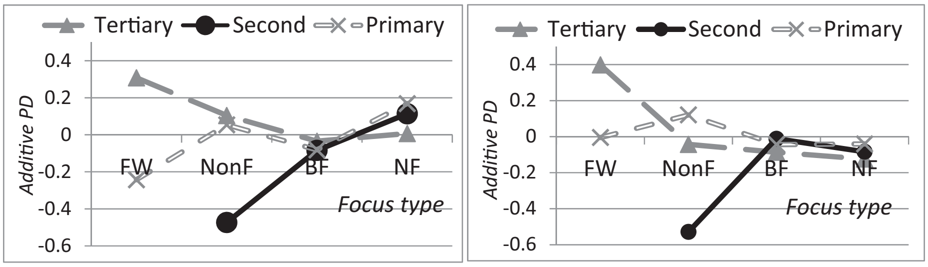

Fig. 12. L1 (Left) and L2 (Right) duration patterns by focus degree after subtraction of segmental, stress and boundary effects. Function word, non-focus, broad focus and narrow focus are labeled FW, NonF, BF, and NF respectively.

shortened. As a result, at the NF position, additive duration incremental on canonical primary/secondary stress surpassed tertiary stress. In other words, interaction with NF position resulted in L1's further lengthened primary/secondary stress and shortened tertiary stress, as shown in the left panel of Fig. 10, and resulted in increased contrast degree among stress types. However, patterns from L2 data exhibited equal duration incremental among primary, secondary and tertiary at the NF position, as shown in the right panel of Fig. 12. The results demonstrate that lower discrimination among stress types in L2 patterns; results shown in Fig. 10 retains.

5.2.5. *Summary of temporal differences*

The duration study showed how Taiwan L2's tempo contrast is less distinct than L1 speech in relation to each specification and from lower to higher levels, and accounted for L2 accent due to temporal patterns. Patterns of temporal adjustment at the segmental level suggest that L2 speakers' greatest challenge lies in the production of central vowels, back vowels and fricatives. At the level of lexical stress, the same L2 speakers exhibited a lower degree of contrast between stressed and unstressed syllables in all conditions, an echo to our f0 findings and another account of L2 accent. The temporal contrasts used to mark different levels of prosodic boundaries were also less strongly differentiated in L2 speech, as found in pre-boundary lengthening/shortening contrasts across non-phrase boundaries, continuation rises, final rises and final falls. As for the temporal adjustments used to mark focus status, while L1 speakers synchronized their temporal adjustments for focus status with lexical-stress specifications. As a result of interaction, temporal contrasts of lexical stress increased with increasing levels of focus. However, this interaction caused adjustment was not found in L2 speech. Taken together, these results suggest that encoding higher-level prosodic information, such as boundary and focus marking, presents the great challenge to L2 speakers. The duration results at both lower and higher levels also allowed us to compare the L1-L2 difference in a hierarchical/additive/finer-grained perspective which better accounts for L2 prosodic accent, and distinguished our present to most previous L2 studies which investigated individual level without considering layered interactions.

5.3. *L1 prosody prediction using linguistic specifications*

5.3.1. *Objective evaluation of F0 prediction*

Predicted F0 and tempo using our L1 prosody model (see Section 4) were objectively evaluated with features derived from baseline model according to RMSE and Corr. Contribution weights among each linguistic/prosodic level in our linear model were further examined to account for prosodic spectrum of native/expressive L1 prosody

5.3.1.2. Modeling position of accent command. Using a decision tree to predict position of accent command, this model achieved an overall prediction accuracy rate of 93.66% (Su et al., 2016a). The strongest predictors by linguistic specifications are 'contrast with previous focus', 'focus degree', 'focus structure', 'information density by syllable', and 'information density by word'.

5.3.1.3. Modeling Aa. RMSE and Corr between generated and original Aa values by different methods are presented in Fig. 13. Overall, the current model performed better than the baseline, with an average improvement over three regression analyses (MLR, RoFit and FNN) of 0.16 (baseline) to 0.13 (proposed model) in RMSE, and 0.45 to

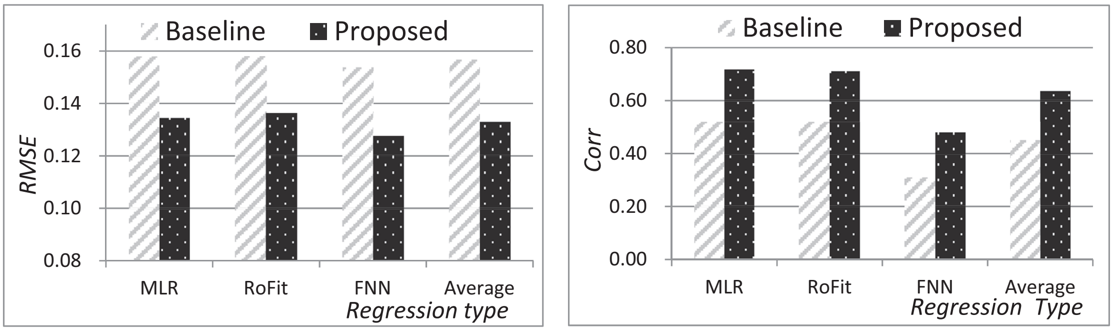

Fig. 13. RMSE (Left) and Corr (Right) between predicted Aa and original Aa extracted using multivariable linear regression (MLR), Robust regression (RoFit) feedforward neural network (FNN).

0.64 in Corr. The most substantial Corr and RMSE improvements of our model are found in performance over the baseline, as demonstrated in MLR analysis (RMSE: 0.16 to 0.13; Corr: 0.52 to 0.72).

Results of weight contribution analyses using MLR and RoFit are given in Table 5 (Su et al., 2016a). The top three contributing linguistic specifications across both regression analyses were 'focus degree of current word', 'Type of lexical stress in current syllable' and 'relative syllable position by primary stress'.

5.3.1.4. Modeling Ap. RMSE and Corr between generated and original Ap values by different methods are presented in Fig. 14. Three RMSE analyses determined that average improvement showed how our model performed over the baseline, that is, 0.19 (baseline) to 0.18 (current), with the most significant RMSE improvement found in FNN (0.21 to 0.17). Average improvement measured by three Corr analyses was 0.34 (baseline) to 0.48 (proposed), with the most significant improvement found in FNN (0.37 to 0.62)

Table 6 presents results of MLR and Rofit weight contribution analyses (Su et al., 2016b), both of which showed that the three strongest linguistic predictors are 'Distance to pre Ap', 'Normalized position by BG' and 'Normalized position by PG'. Information density', which correlates most positively to information content, was ranked the fifth.

5.3.1.5. Summary of objective evaluation of F0 prediction for L1. The F0 results demonstrate that the proposed method can improve the baseline model across different levels of features and different types of regression models. The Aa position and magnitude at the lower level predicted using the proposed method suggests that, for L1 speakers, the location of local F0 humps is systematic and predictable using a combination of linguistic information related to lexical stress, focus structure and discourse structure. The three strongest predictors are focus structure > lexical stress > information density. Comparing different regression models for Aa prediction (magnitude of short-term/local F0 humps) in the proposed method, an interesting finding is that simpler linear combinations

Table 5
Relative contribution weights or Aa prediction.

Regression\Contextual linguistic specification	MLR	RoFit
X1 - Lexical stress type of current syllable	2.36	2.20
X2 - Pre-syllable contrast within a word	−1.89	−1.74
X3 - Post-syllable contrast within a word	0.09	0.15
X4 - Relative syllable position by primary stress in a word	2.41	2.30
X5 - Focus degree of current word	2.88	2.71
X6 - Pre-word contrast within a sentence	0.09	0.02
X7 - Post-word contrast within a sentence	0.58	0.63
X8 - Relative position to NF in a sentence	0.99	0.36
X9 - Relative position to first BF in a sentence	−1.03	−0.74
X10 - Boundary type (level) of current phrase	−1.56	−1.97
X11 - Boundary type (level) of pre phrase	1.38	1.74
X12 - Boundary type (level) of post phrase	−0.89	−0.92
X13 - Infodensity by syllable	−1.03	−0.48
X14 - Infodensity by word	1.61	1.52

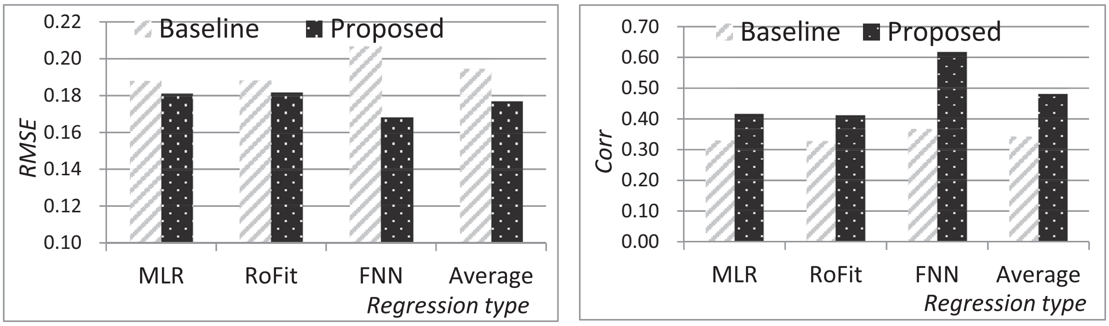

Fig. 14. RMSE (Left) and Corr (Right) between predicted Ap and original Ap extracted by MLR, RoFit and FNN.

outperformed more complex model of multi-layer nonlinear transformation, especially in terms of correlation. This suggests the production/planning of short-term/local F0 humps is a straightforward scheme which only requires simple linear combination. As for Ap (global F0 magnitude) prediction at higher level using the proposed method, contributing weight analyses demonstrate that between-Ap/PPh association, normalized BG and PG positions are the three strongest predictors, which suggests that the patterns in global F0 magnitude produced by L1 speakers are strongly influenced by discourse (and thus hierarchical) structure. The comparison across regression models for Ap prediction using the proposed method demonstrated the non-linear model outperformed the linear model, suggesting the L1 F0 production of long term/global discourse units at higher level is a relatively more cognition-intensive process than linear combination. As a result, the prediction for local and global F0 features of L1 allowed us to examine improvements our proposed features and comparison with the baseline models captured, as well as facilitated better understanding of F0 production/planning of native/expressive L1 prosody using machine simulations.

5.3.2. Objective evaluation of tempo prediction

Fig. 15 shows a comparison of RMSE and Corr analyses of phoneme duration (PD) predicted by the baseline and the current model. Overall, the current model performed slightly better than the baseline. An average of three RMSE analyses showed an improvement of 0.55 (baseline) to 0.53 (proposed), as well as the MLR analysis, which showed and improvement of 0.51 to 0.49. Average improvement shown in three Corr analyses was 0.72 (baseline) to 0.73 (proposed); FNN analysis results showed an improvement from 0.78 to 0.79. Subsequent analyses of contribution weights using MLR and RoFit appear in Table 7 (Su et al., 2016c).

5.3.2.2. Summary of L1 tempo prediction results. Overall, the results suggest that for L1 speakers, PD is systematically determined through a combination of linguistic specifications related to phoneme identity, stress type, syntactic structure and focus degree while the present method slightly improved prediction accuracy than the baseline model. However, we believe the predictions for temporal features allowed us to better understand tempo production/planning of native/expressive L1 speech using machine simulation.

Table 6
Relative contributing weights for Ap prediction.

Higher level specifications\Regression		MLR	RoFit
Discourse structure	Phrase length by word number	1.37	0.66
	Break level preceding a phrase	0.53	0.53
	Break level following a phrase	0.10	0.09
	Normalized phrase position in a PG	2.43	3.63
	Normalized phrase position in a BG	6.17	7.11
	Distance to pre phrase by word number	11.37	10.88
Information allocation	Number of BF and NF per phrase	−1.90	−1.20
	Number of NF per phrase	0.00	0.05
	NF position in a phrase	0.88	0.45
	Information density in a phrase	1.44	0.98

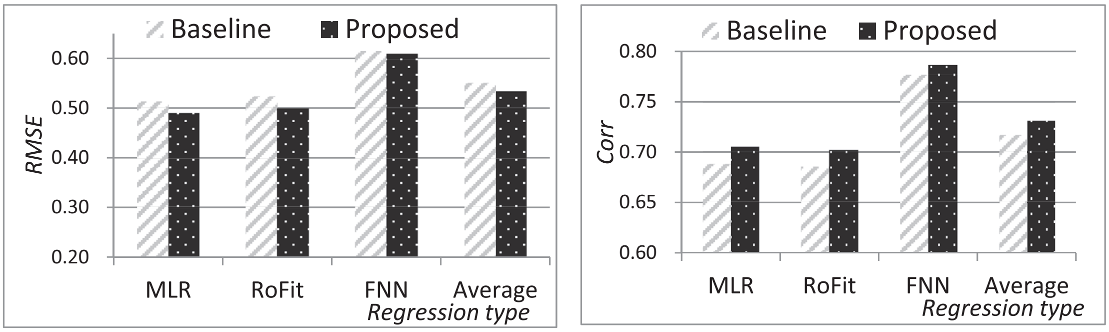

Fig. 15. RMSE (Left) and Corr (Right) between predicted PD and original PD extracted using MLR, RoFit, and FNN.

Table 7
Relative contributing weights for tempo prediction.

Linguistic specification\Regression	MLR	RoFit
phoneme identity of current phoneme	0.97	0.99
lexical stress of current syllable	1.04	0.85
focus degree of current word	1.08	0.93
syntactic structure of current phrase	0.97	0.54

5.3.3. Subjective evaluation of F0 and tempo predictions

The current L1 prosodic model was used to resynthesize L2 speech, which was subsequently compared with original L2 speech tokens in order to determine which speech samples (original or resynthesized) sounded more ‘native-like’ . Results of a perception test using native listeners are presented in Fig. 16. The dotted, meshed and twilled partitions in each bar respectively represent the percentage of resynthesized L2 speech perceived as (1) improved (2) worsened (3) no difference from the original L2 speech by perceived accent. The largest partitions at each prosodic level, marked by * in Fig. 16 demonstrate that resynthesized versions (red) are perceived as most ‘native-like’ at each level. The results therefore suggest L2 speech with our modifications did reduce the degree of accent in comparison to the original L2 speech by each specified prosodic level.

At the lexical level, L2 speech that had been resynthesized using our F0 and duration model received the highest rating (82.14% of listeners chose these tokens). At the sentence level, L2 speech resynthesized with F0 and F0 + Dur models received ratings of 75% and 58.3%, respectively, both of which surpassed the Dur only modele. At the discourse level, speech resynthesized with the Dur model received rating of 62.5% which was also higher than the original L2 speech.

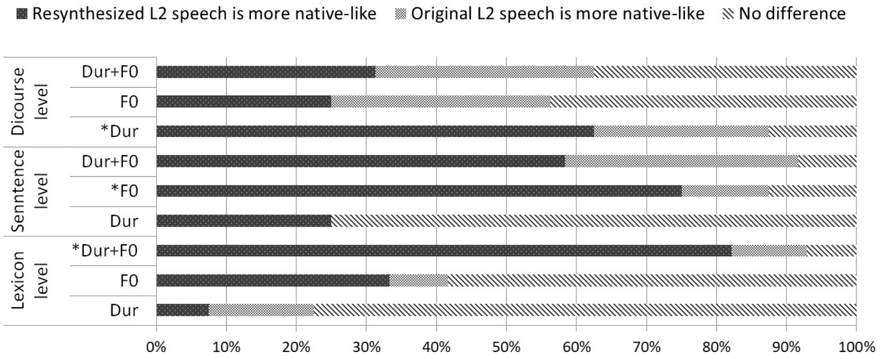

Fig. 16. Perception results.

5.3.4. *Summary of L1 prosody prediction and evaluation*

The results suggest that resynthesizing L2 speech using the prosodic parameters generated by the current model substantially reduced the level of perceived L2 accent at each level of specific prosodic features. However, at the discourse level, improvement of perceived L2 accent using the F0 and Dur + F0 models is greater than the Dur only model. Our speculation is such that this may be due to the complex F0 generation processes from automatic extraction of Ap/Aa to curve generation as well as from lower (syllable level) to higher level (discourse/information structure). The procedures may also involve configuration of particular parameters not yet addressed in the present study such as onset time of phrase component, T_{0i} in Eq. (1), which is set to a fixed value for the time being in the present study. In other words, perhaps some inappropriate parameters configured from lower level or even from the extraction procedure may cumulate to the higher level whereby accumulated errors may be the cause. Nevertheless, this phenomenon requires further studies and more refined parameter adjustments at each level for confirmation.

6. General discussion

The present study is the first attempt to incorporate both hierarchical discourse and focus-type defined information structure to (1) investigate the interactive contributions of prosodic differences between L1 and L2 as well as to (2) simulate native/expressive L1 prosody for generating corrective feedback to improve L2 expressive prosody. Using lower- to higher-level specifications related to linguistic, discourse and information structures, with special attention to help enhance semantic association/cohesion and weighted information placements, the present study features L1/L2 comparison of contrast differentiation across a wider range of features than most reported CALL systems set goals on, and targets to bring learners' attention to the prosodic composition and characteristics that contribute derivationally to fluency and expressiveness. The proposed range of linguistic specifications spans from lower to higher levels; their interactive contribution to output prosody clearly manifested. That is, at the lower (word and sentence) levels, L2's weaker contrast differentiation, substantiated by contrastive patterns of F0 and tempo, is identified to relate to word stress patterns and focus status. At the higher (multi-phrase discourse and information) levels, L2's weaker contrast differentiation is also identified to cross-phrase cohesion and higher level foci. As a result, when lower level contrasts are further nested into higher and larger units, their interaction causes lower level contrasts to further weaken and the higher level contrasts less obvious. Our results also suggest that when producing continuous speech, Taiwanese L2 English speakers tend to plan by smaller speech units (words and single sentences, Tseng et al., 2010), did pay attention to lower level features (word stress and sentence intonation), but are less able to achieve the required degree of contrast differentiation. At the same time, they (Taiwanese L2 speakers) are less sensitive to other multiple levels of prosodic planning involved to produce continuous speech, most notably cross-phrase cohesion from a hierarchical perspective and marking information structure. The interactive results have been accounted for.

Using the extracted patterns/modules of L1 prosody at each linguistic/prosodic level from low to high, the present study proposes a prosodic model which aims to simulate native/expressive L1 speech for generating corrective prosodic feedback for L2 learners. When compared to a baseline model the proposed model achieved a higher level of accuracy in prediction and regeneration of L1's F0 and duration, in which (1) input variables are directly extracted from a flat level of surface prosody without separating mutual interaction between layers before modeling, and (2) discourse and information structures are not added to the list of predictor variables. The relative improvement using our proposed features and method provided better simulation results which in turn also accounted for production/planning of native/expressive L1 prosody.

Using our methods of simulating the production of L1 prosody, contribution weights for each linguistic/prosodic level with a linear model are compared across Aa, Ap and PD to account for the prosodic spectrum of native/expressive L1 speech. The comparison shows that the novel features "information density" and "discourse specifications" do contribute to surface out prosody; their respective contributions different. Regarding short-term/local F0 humps (Aa), information-related specifications such as local and nearby information contents play a more important role than discourse-related specifications such as boundary types. We believe this is a special feature of continuous speech planning that merits more research attention. Regarding long term/global tendency of F0 (Ap), discourse structures such as between-Ap association, BG and PG positions appear to dominate the contributions of Ap prediction. As for duration features (PD), discourse and information structures contribute evenly to the final output (surface form). The results account for L1's complex strategy involved in planning multiple prosodic levels by different sizes with aims to express semantic association/cohesion and highlight key information through prosody specific to fluent continuous speech. Needless to say, the task is difficult for L2 learners and explicit a corrective feedback model should be helpful.

Comparison of the linear model and multi-layer non-linear transformation for F0 modeling of L1 yielded different performance results between two levels for F0 features, namely, local F0 humps (Aa) and global F0 tendency (Ap), thus nailing down different degree of planning complexity for respective corresponding levels. Prediction of local F0 humps (Aa) showed the simpler linear model outperforms non-linear transformation whereas prediction of global F0 tendency (Ap) exhibited opposite results in which complex non-linear model performed better. The simulation results imply that production of long term/global F0 for discourse units involves more complex planning processes than short-term/local units, thus further accounts for L2 speakers' difficulty when trying to plan long-term/global features, as similarly evidenced in previous works of how F0 features related to both discourse and information structure (Atterer and Ladd, 2004; O'Brien and Gut, 2010; Nguyễn et al., 2008; Saha and Mandal, 2017; Tseng et al., 2010).

It is therefore our belief that when used to improve the quality of user feedback in CALL applications, our findings could help draw L2 speakers' attention to the range of higher-level prosodic features involved in production of fluent, expressive speech and used to increase the efficiency of L2 prosody training paradigms. Resynthesized L2 speech using our model was also tested for subjective perceptual support. The results demonstrated that the current model substantially reduced the level of perceived accent in L2 speech. If incorporated into a CALL system, the proposed model is able to generate prosodic feedback for L2 learners by predicting L1 prosodic contours from annotated text with at least three advantages: (1) most linguistic specifications and prosodic features used for modeling have been developed in studies across languages and suggest these specifications are high-level/additive features upon lower-level intrinsic properties such as segments/speech contents, etc. In other words, even when lower-level speech contents are not available in the L1 database, the L1 prosodic patterns can still be predicted to simulate fluent and expressive prosody as corrective norm as long as higher-level linguistic information above segments is provided or labeled on the text. (2) Since corrective prosody was superimposed on speech segments produced by L2 leaners themselves, hearing the modified prosodic outcome of their own voices should bring more pronounced awareness of the L1/L2 differences. (3) By breaking down the relative contributions from different levels of specifications, the proposed model is capable of providing customized adaptive feedbacks to identify particular difficulties of each leaner by different levels, for example using F0 + Dur model to train better production of lexical stress.

Moreover, existing TTS systems such as HTS (HMM-based speech synthesis system, Tokuda et al., 2002) reports that output naturalness is achieved, but at the cost of expressiveness due to over-smoothing. We believe systems as such could benefit from incorporating our model to produce more pronounced discourse-association and more elicited information foci. Therefore, another possible extension of application is to integrate our model into existing TTS/CALL applications to further enhance the comprehensibility and expressiveness of output prosody due to discourse cohesion and information structure.

7. Conclusion

The proposed model allows for the more accurate characterization of the differences between the prosody of L1 and Taiwan L2 English speech with a hierarchical account of global discourse factors as well as information structure, thereby incorporating interaction of multiple levels of interaction involved in overall continuous speech prosody. Experimental results revealed how Taiwan L2 speakers are less sensitive to different sized speech units and hence supported our motivation that discourse cohesion, information structure and their interaction were indeed the major contributors to Taiwan L2 English accent. Specific difficulties Taiwan L2 English speakers experience in producing fluent and expressive prosody was further accounted for by analysis and simulation of L1 prosody, where complex parallel planning strategies to achieve semantic association/cohesion and emphasize key information contents at the same time are involved. Our L1 prosodic model was used to resynthesize corrected L2 speech to simulate improved prosody for L2 learners with feedbacks that help increase both intelligibility and comprehensibility in terms of fluency and expressiveness. We hope that the proposed linguistic information-based prosody model could further enhance fluency, comprehensibility and expressiveness of L2 continuous speech and at the same time provide more linguistic implications for advanced computer-assisted language learning. Future research will focus on training the proposed model with more realistic speech data to improve its prediction accuracy, with particular attention to more refined parameters and adjustments at the discourse level, and its integration with existing TTS/CALL systems.

Acknowledgment

This research is partially supported by The Ministry of Science and Technology, ROC (Grant no. MOST 104-2221-E-002-051-MY).

Appendix A. Target words by syllabicity, stress type and experimental condition.

Stress type	Target words	Stress type	Target words
2−1	**Mo**ney; **mo**rning	4−1	**E**levator; **Ja**nuary
3−1	**Vi**deo; **ho**spital	4−2	A**vail**able; ex**pe**rience
3−2	A**part**ment; to**mo**rrow	4−3	Infor**ma**tion; Cali**for**nia
3−3	Overnight; Japanese	4−4	Misunderstand; Vietnamese
Left-Headed compound	**su**permarket; de**part**ment	Right-Headed compound	White **wine**; after**noon**

Where 2−1 represents two-syllable words with primary stress in 1st syllable position

Appendix B. Carrier sentences.

Examples

1 I said **EXPERIENCE**$_{\textit{(TARGET WORD)}}$ ten times.
2 I said **VIDEO**$_{\textit{(TARGET WORD)}}$ five times.
3 I said **ELEVATOR** $_{\textit{(TARGET WORD)}}$ five times.

Appendix C. At prosodic boundaries.

An Example

Although Fred didn't have any **EXPERIENCE**$|_{\boldsymbol{CR}}$, he had no trouble learning how to make a **VIDEO**$|_{\boldsymbol{FF}}$.

Where CR: continuation rise, FF: final fall, the other words without label: Non-phrase boundary.

Appendix D. In the position of contrastive stress.

Examples

1 Context: Have you been trained to do this job?

Answer: No. But I think **EXPERIENCE**$_{(NF)}$ is more important$_{(BF)}$ than training$_{(BF)}$.

1 Context: Are we allowed to make audio and video recordings?

Answer: No. **VIDEO**$_{(NF)}$ recordings$_{(BF)}$ are not$_{(BF)}$ allowed$_{(BF)}$.

1 Context: How will I carry all these boxes up to the fifth floor?

Answer: You should take$_{(BF)}$ the **ELEVATOR**$_{(NF)}$ instead of the stairs.

Where NF: narrow focus, BF: broad focus, the other words without label: Non-focus.

References

Andersen, R., "Modern methods for robust regression", Sage University, Paper Series on Quantitative Applications in the Social Sciences, 2008.

Auer, P., Harald, B., Wolfgang, M., 2008. A learning rule for very simple universal approximators consisting of a single layer of perceptrons. Neural Netw. 21.

Anderson-Hsieh, J., Johnson, R., Koehler, K., 1992. The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. Lang. Learn. 42, 529–555.

Atterer, M., Ladd, D.R., 2004. On the phonetics and phonology of ‘‘segmental anchoring’’ of F0: evidence from German. J. Phonet.

Andreeva, B., Barry, W.J., Koreman, J., 2016. Local and global cues in the prosodic realization of broad and narrow focus in Bulgarian. Phonetica 73 (3-4), 256–278.

Avesani, C., Vayra, M., 2003. Broad, narrow and contrastive focus in Florentine Italian. In: Proceedings of the 15th International Congress of Phonetic Sciences, 2, pp. 1803–1806.

Bailly, G., Holm, B., 2005. SFC: a trainable prosodic model. Speech Commun. 46, 348–364.

Benrabah, M., 1997. Word-stress: A source of unintelligibility in English. IRAL XXXV (3), 157–165.

Coniam, D., 1999. Voice recognition software accuracy with second language speakers of english. System 27 (1), 49–64.

Chafe, WL., 1974. Language and consciousness. Language 1974 (50), 111–133.

Derwing, T.M., Munro, M.J., 1997. Accent, intelligibility, and comprehensibility: evidence from four L1s. Stud. Second Lang. Acquis. 19 (1), 1–16.

Domínguez, M., Farrús, M., Burga, A., Wanner, L., 2014. The Information structure-prosody interface revisited. In: Proceedings of 7th International Conference on Speech Prosody, Dublin, Ireland.

Domínguez, M., Farrús, M., Burga, A., Wanner, L., 2016. Using hierarchical information structure for prosody prediction in content-to-speech applications. In: Proceedings of the 8th International Conference on Speech Prosody, Boston, USA.

Fujisaki, H., 2004. Information, prosody, and modeling-with emphasis on tonal features of speech. In: Proceeding of the Speech Prosody, Nara, Japan.

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 Commun. 47, 59–70.

Gananathan, R.Y., Yin, Y., Ki, K., Mok, P., 2015. Interlanguage influence in cues of narrow focus: a study of Hong Kong english. In: Proceedings of the International Conference on Phonetic Sciences 2015, Glasgow UK. August.

Grosser, W., 1997. On the acquisition of tonal and accentual features of English by Austrian learners. In: James, A., Leather, J. (Eds.), Second Language Speech: Structure and Process. De Gruyter, Berlin, pp. 211–228.

Gut, U., 2009. Non-Native Speech: A Corpus-Based Analysis of Phonological and Phonetic Properties of L2 English and German. Frankfurt. Peter Lang, Germany.

Gut, U., Pillai, S., Mohd, D.Z., 2013. The prosodic marking of information status in Malaysian English. World Englishes 32 (2), 185–197.

Handley, Z., 2009. Is text-to-speech synthesis ready for use in computer-assisted language learning? Speech Commun. 51 (10), 906–919.

Hirose, K., Fujisaki, H., Yamaguchi, M., 1984. Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information. In: Proceedings of the ICASSP.

Halliday, MAK., 1967. Notes on Transitivity and Theme in English, Part 2. J. Ling. 1967 (3), 199–244.

Hoskins, S.R. The prosody of broad and narrow focus within noun phrases (Doctoral dissertation, ASA).,1997

Hanssen, J.E.G., Peters, J., Gussenhoven, C., 2008. Prosodic effects of focus in Dutch declaratives. In: Proceedings of Speech Prosody 2008.

James, E., Atkinson, J., 1976. Inter− and intraspeaker variability in fundamental voice frequency. Acoust. Soc. Am. 60 (2).

Jacewicz, E., Fox, R.A., Wei, L., 2010. Between-speaker and within-speaker variation in speech tempo of American English. J. Acoust. Soc. Am. 128 (2).

König, E., 2002. The Meaning of Focus particles: A comparative Perspective. Routledge.

Laver, J., 1991. The Gift of Speech. Edinburgh University Press, Edinburgh, UK.

Lambrecht, K., 1994. Information Structure and Sentence Form. Cambridge University Press, Cambridge, UK.

Moustroufas, N., Digalakis, V., 2007. Automatic pronunciation evaluation of foreign speakers using unknown text. Comput. Speech Lang. 21 (1), 219–230.

Munro, M.J., 1995. Nonsegmental factors in foreign accent: ratings of filtered speech. Stud. Second Lang. Acquis. 17, 17–34.

Mixdorff, H., 2000. A novel approach to the fully automatic extraction of fujisaki model parameters. In: Proceedings of ICASSP 2000. 3, Istanbul, Turkey, pp. 1281–1284.

Mixdorff, H., 2002a. Speech Technology, ToBI and Making Sense of Prosody. Invited talk at Speech Prosody 2002. Aix, France, pp. 31–38.

Mixdorff, H, 2002b. An integrated approach to modeling german prosody. Studientexte zur Sprachkommunikation, 25, Dresden.

Malah, D., 1979. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. IEEE Trans. Acoustics, Speech Signal Process. ASSP-27 (2), 121–133.

Nakamura, S., 2010. Analysis of relationship between duration characteristics and subjective evaluation of english speech by japanese learners with regard to contrast of the stressed to the unstressed. J. Pan-Pacific Assoc. Appl. Ling. 14 (1), 1–14.

Nespor, M., Vogel, I., 1986. Prosodic phonology. Mouton de Gruyter, Dordrecht: Foris, Berlin.

Nguyễn, T.A., Ingram, C.L.J, Pensalfini, J.R., 2008. Prosodic transfer in Vietnamese acquisition of English contrastive stress patterns. J. Phonetics. 2008 36 (1), 158–190.

O’Brien, M., Gut, U., 2010. Phonological and phonetic realisation of different types of focus in L2 speech. In: Kul, M. (Ed.), Dziubalska-Kołaczyk Katarczyna, Wrembel Magdalena. Peter Lang, Frankfurt, pp. 205–215. Achievements and perspectives in the acquisition of second language speech: New Sounds 2010.

Pedhazur, EJ., 1982. Multiple Regression in Behavioral research: Explanation and Prediction. 2nd ed. Holt, Rinehart and Winston, New York.

Perwitasari, A., Klamer, M., Witteman, J., Schiller, N.O., 2015. Vowel duration in English as a second language among Javanese learners. In: Proceedings of the International Conference on Phonetic Sciences 2015, Glasgow UK. August.

Peppé S., Maxim J., Wells B. Prosodic Variation in Southern British English Language and Speech, Language and speech, 2000.

Ramirez Verdugo, M.D., 2002. Non-native interlanguage intonation systems: a study based on a computerized corpus of Spanish learners of English. ICAME J. 26, 115–132.

Scruton, R., 1996. The Eclipse of Listening, 15. The New Criterion, pp. 5–13.

Selkirk, E., 1984. Prosody and syntax: The relation between sound and structure. MIT Press, Cambridge, MA.

Su, C.Y., Tseng, C.Y., 2016a. L1/L2 difference in phonological sensitivity and information planning - evidence from F0 patterns. In: Proceedings of the ISCSLP2016, Tianjin, China.

Su, C.Y., Tseng, C.Y., 2016b. Global F0 features of mandarin L2 english - reflection of higher level planning difficulties from discourse association and information structure. In: Proceedings of the Oriental-COCOSDA 2016, Bali, Indonesia.

Su, C.Y., Tseng, C.Y., 2016c. The long road from phonological knowledge to phonetic realization − an acoustic account of the temporal composition of mandarin L2 english. Speech Prosody 2016, 16–20.

Sityaev, D., House, J., 2003. Phonetic and phonological correlates of broad, narrow and contrastive focus in English. In: Proceedings of the 15th ICPhS. (Vol. 1822).

Saha, S.N., Mandal, S.K.D., 2017. Discourse prosody planning in native (L1) and nonnative (L2) (L1-Bengali) English: a comparative study. Int. J. Speech Technol. 20 (2), 305–326.

Trofimovich, P., Baker, W., 2006. Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. Stud. Second Lang. Acquis. 28, 1–30.

Tokuda, K., Zen, H., Black, A.W., 2002. An HMM-based speech synthesis system applied to English. In: Proceedings of the 2002 IEEE SSW. Sept.

Tseng, C.Y., Pin, S.H., Lee, Y.L., Wang, H.M., Chen, Y.C., 2005. Fluent speech prosody: framework and modeling. Speech Commun. 46 (34), 284–309 Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation.

Tseng, C.Y., Su, Z.Y., 2008. What’s in the F0 of Mandarin Speech −Tone, Intonation and beyond. In: Proceedings of the ISCSLP 2008. Kunming, China, pp. 45–48.

Tseng, C.Y., Su, Z.Y., Huang, C.F., Visceglia, T., 2010. An initial investigation of L1 and L2 discourse speech planning in English. In: Proceedings of the 2010 7th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, pp. 55–59.

Tseng, C.Y., Su, C.Y., Visceglia, T., 2013. Underdifferentiation of English Lexical Stress Contrasts by L2 Taiwan Speakers. In: Proceedings of the Slate 2013. Grenoble, France, pp. 164–167.

Tseng, C.Y., Su, C.Y., 2014. Prosodic differences between taiwanese L2 and North American L1 speakers — under-differentiation of lexical stress. In: Proceedings of the 7th Speech Prosody Conference, Dublin, Ireland.

Thorén, B., 2007. Swedish accent - duration of post-vocalic consonants in native swedes speaking English and German. In: Proceedings of the International Conference on Phonetic Sciences 2007, Saarbrücken Germany. August.

Witt, S.M., Young, S.J., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Commun. 30 (2-3), 95–108.

Wennerstrom, A., 1994. Intonational meaning in English discourse: A study of non-native speakers. Appl. Ling. 15, 399–420.

Utgoff, P.E., 1989. Incremental induction of decision trees. Machine learning 4 (2), 161–186.

Visceglia, T., Tseng, C.Y., Kondo, M., Meng, H., Sagisaki, Y., 2009. Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project). In: Proceedings of the Oriental COCOSDA 2009, Beijing, China.

Visceglia, T., Su, C.Y., Tseng, C.Y., 2012. Comparison of English Narrow Focus Production by L1 English, Beijing and Taiwan Mandarin Speakers. In: Proceedings of the Oriental COCOSDA. Macau, China, pp. 47–51.

Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. Speech Commun. 46, 220–251.

Zellner, K.B. and Keller, E. "representing speech rhythm", Improvement in Speech Synthesis 154-164, 2001.