Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

An Analysis-by-Synthesis Study of Mandarin Speech Prosody

Na Zhi [1, 2], *Daniel Hirst* [3], *Pier Marco Bertinetto*[4], *Aijun Li* [2], *Yuan Jia*[2]

[1] Capital Normal University, Beijing, China
[2] Chinese Academy of Social Sciences, Beijing, China
[3] Aix-Marseille University, CNRS, Aix-en-Provence, France
[4] Scuola Normale Superiore di Pisa, Italy

zhinacn@yeah.net, djhirst@me.com, piermarco.bertinetto@sns.it, liaj@cass.org.cn, summeryuan_2003@126.com

Abstract

In the present paper an analysis by synthesis study of mandarin speech prosody is carried out. The mandarin prosodic features are discussed from two salient perspectives, specifically: the function of prosody and the form of prosody. The symbolic representation of prosodic form with the INTSINT (INternational Transcription System for INTonation) system [1] reduces the surface complexity of a prosodic contour to a simplified model, which contains the essential information expressing the functions of speech prosody. A proposed mapping rule between the representation of prosodic function and the representation of prosodic form is discussed and further evaluated in ProZed [2, 3, 4, 5] by generating synthesized utterances. It is suggested in the study that the synthesized mandarin data derived from the prosodic coding of INTSINT symbols can not only closely mirror the melodic features of the original utterances, but also correctly express the prosodic functions of tones and the global intonation.

Index Terms: analysis, synthesis, mandarin, INTSINT system, ProZed

1. Introduction

How to relate the physical acoustic information of the speech form in an appropriate way with its specified prosodic function is still a poorly understood problem in analyzing and modelling the prosody of natural languages. To provide a potential solution to the mapping between the function and the form of speech prosody, a multi-level organization for the form-function interface is postulated in [6, 7]. By means of an "analysis-by-synthesis" paradigm, linguists are encouraged to define mapping rules between the formal and functional aspects of prosody [3], as revealed in Figure 1.

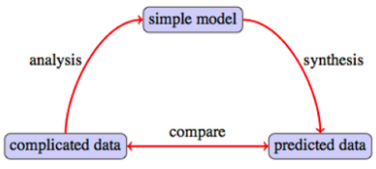

Figure 1: *The "analysis-by-synthesis" paradigm of speech production.*

The above paradigm shows a cyclic process in which the raw F0 contour (as the "complicated data") can be analyzed and annotated within a "simple model" with the help of prosodic coding symbols, e.g. the INTSINT alphabets [1]. The resulted symbolic representation of prosody can then be used for generating synthesized utterances in ProZed [2, 3, 4, 5] for perceptual evaluation; the acoustic data "predicted" from the prosodic analysis can then be compared to the original raw F0 contour for evaluation. Such a cyclic implementation allows linguists to test and evaluate their theoretical proposals directly with the resulted acoustic signals owing to the development of speech synthesis technology.

In the present study, we follow the above process in studying mandarin speech prosody. By analyzing the functional information of mandarin tones and intonation, as well as by coding the physical facts of mandarin prosodic form, we derive a simplified model with INTSINT symbols stylizing the melodic events. Such a symbolic representation is inut to the ProZed tool for speech synthesis. The synthesized data is examined to see whether it can closely capture the pitch features of the utterance at the physical level, and whether it contains all the necessary information, both lexical and sentential, for expressing the functions of mandarin prosody. The goal is to find a plausible way to model the functional and formal annotation and analysis of mandarin speech prosody.

2. INTSINT-Momel system and ProZed

For the symbolic coding and automatic modelling of prosodic patterns, we firstly introduce two systems which are implemented in ProZed. They are the INTSINT coding system and the Momel (MOdelling MELody) algorithm [8].

The INTSINT system employs an alphabet of 8 letters for annotating the surface pitch movement. They are t(op), m(iddle), b(ottom), h(igher), s(ame), l(ower), u(pstepped) and d(ownstepped). The points *t*, *m*, and *b* are **absolute tones**, corresponding to the relevant positions in a speaker's pitch range (defined with two parameters, key and span); the tones *h*, *l*, *s*, *u* and *d* are **relative tones**, each defined with respect to the immediately preceding tonal target. Figure 2 illustrates the relations of INTSINT targets from [9]:

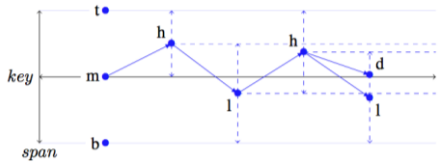

Figure 2: *The INTSINT alphabets from [9].*

The INTSINT labels can be used as the direct input for the Momel algorithm, which interpolates the target points in sequence with a quadratic spline function, resulting in transitions between smooth and continuous points. A sample

10.21437/SpeechProsody.2016-22

of the modelling output with the Momel algorithm can be seen in the following Figure 3 taken from [10]:

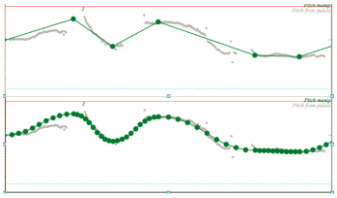

Figure 3: *The modelling output with the Momel algorithm [from 10]*

Figure 3 shows that, with the help of the INTSINT-Momel system, a raw F0 contour can be reduced to and defined by a number of salient pitch targets, which are enough to capture the melodic movement. The two systems are now integratedin a speech tool, ProZed which has beenimplemented as a plugin to the Praat program [11] and easily be accessed byresearchers to test their analysis-by-synthesis paradigm of speech prosody.

Although the above system was not originally designed for the analysis of a tone language, one can assume that the melodic properties of Mandarin speech can be studied in the same way as those of non-tonal languages by reducing the observable complexity stemming from a large quantity of data to a more simplified model, preserving all the necessary information for expressing the functional contrasts in speech. In Mandarin speech, the interactive relation between lexical tones and the "purely intonational use" [12] of pitch prosody makes the whole F0 picture very intricate, so as to make it of the highest interest for prosodic studies. In the following section, we present the functional and formal analysis of prosodic events in Mandarin.

3. Events of Mandarin speech prosody

The Mandarin speech melody is a combined carrier of pitch information at both the lexical level and the sentential level. Therefore, prosodic events in Mandarin can be decomposed into two levels of functional representation. At the lexical level, the tonal contour contributes to the specification of word identity; at the sentential level, intonational prosody contributes to marking sentential prominence and defining prosodic boundaries, namely, the "weighting function" and the "grouping function" of intonation in [13]. The functional annotation of the prosody of a pragmatically neutral Mandarin utterance could be presented as follows:

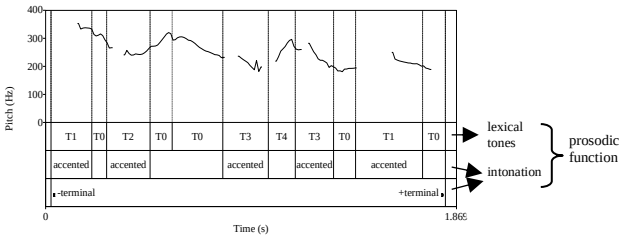

Figure 4: *The representation of prosodic function of a neutral Mandarin utterance.*

3.1. Tonal events in mandarin prosody

In Mandarin the pitch value of each syllable contributes to the lexical distinctive function. There are four distinct tones, namely: tone 1 (high-level tone, HH), tone 2 (rising tone, LH), tone 3 (low-dipping tone, LLH), and tone 4 (falling tone, HL). In addition, standard Chinese exhibits a neutral tone (T ø).

However, in connected speech the tone feature of each syllable can be quite different from its underlying form as produced in isolation. There exists a vast diversity of tone undershoot. The deviant forms of the tones are found when there are no clearly defined boundaries between words, phrases or even sentences in the spoken language. The deviational degrees of the tones are sensitive to factors such as prosodic effect, functional load of the syllables in lexical-syntactic context, speech style, and speech rate [14].

Despite such bewildering surface forms of tone features, it is found that tone sandhi in many cases can be simulated and even predicted by locating the rational domains in which sandhi processes take place. As discussed in [15], tone coarticulation mostly takes place within a certain prosodic-syntactic unit and between "domain partners".

Tonal manifestation is closely related to the accentual status of its aligned syllable in speech [16]. We believe that the surface tonal pattern in connected speech cannot be analyzed at the syllabic level, but rather at a higher level of unit, e.g., the "prosodic unit" in which the tone of a prosodically weak syllables tend to be articulated together with that of the adjacent strong syllable, under the condition that weak syllables syntactically and semantically establish a close relation with the neighboring prominent one.

Detailed exploration of prosodic units and tone dynamics in the speech flow were analyzed with pragmatically neutral utterances, and were listened by 15 native speakers to determine the sentence accents of each utterance [17, 18]. Here we show the result of our analysis using one sample utterance.

Utterance (1): "ta de hai zi men zao jiu deng zhe chi le"

ta	de	hai	zi	men	zao	jiu	deng	zhe	chi	le
She	possessive	child	diminutive	plural	early	already	wait	imperfective	eat	particle

Her children have been long waiting for eating (the cake).

Five syllables in the above utterance were perceived as bearing sentence accent with native subjects' consensus higher than 70%, namely *ta* (90% of agreement), *hai* (80%), *zao* (80%), *deng* (80%), *chi* (90%). The accentual hierarchy of the utterance is represented in the form of a metrical grid with three levels: syllable level, word level and IU (intonation unit) level.

$X_{(90\%)}$		$X_{(80\%)}$			$X_{(80\%)}$		$X_{(80\%)}$		$X_{(90\%)}$		→ IU level
X		X			X	X	X		X		→ word level
X	X	X	X	X	X	X	X	X	X	X	→ syllable level
ta	de	hai	zi	men	zao	jiu	deng	zhe	chi	le	relative prominence

At the syllable level, all audible nuclei within an utterance are indicated by 'X' in the grid representation, assuming that "every syllable participates in the rhythmic organization of the utterance" [19]. At the word level, all content words are marked by 'X' as default strong syllables, while grammatical words are represented by default as unaccented syllables. At the IU level, only the syllables which contribute to the sentence prominence are marked.

At the IU level, accent contributes to chunking speech into hierarchical prosodic units. Such prosodic domains are conceived of as the units where tone sandhi often takes place. As discussed above, tonal realization is closely related to its accentual status in prosody. Within a prosodic unit, the tones of weak syllables undergo tone coarticulation with the adjacent prominent syllable.

The prosodic hierarchy of the sample utterance is displayed in five levels of units, namely: **utterance (U)**, **intonation unit (IU)**, **accentual phrase (Σ)**, **tonal unit (TU)** and **syllable**.

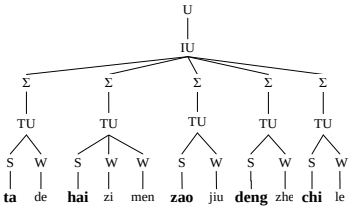

Figure 5: *the prosodic hierarchy of the utterance.*

As can be seen, the five Σs in the utterance are formed by the accented syllables and their following unaccented ones. A TU is defined by both the lexical-syntactic relations of the component syllables and the metrical structure of the unit, as argued in [16]. In the utterance, all Σs coincide with the TUs, as the component syllables of each Σ are close in lexical-syntactic meanings, and meet the formation principle of a TU. The "S" and "W" in the metrical grid correspond to the strong and weak syllables in the utterance.

The citational tonal form of each syllable is represented in the following left figure, where the default neutral-tone syllables (the grammatical particles, *de*, *men*, *zhe*, *le* and the nominal diminutive *zi*) are not marked with tones. In the utterance, syllabic tones undergo a *tone spreading process* [16] according to which a full tone extends its tonal feature rightwards in a TU formed by the accented syllable and the following unaccented one. This causes the citational tone feature of each syllable to span over a larger unit in connected speech, as seen in the following right figure:

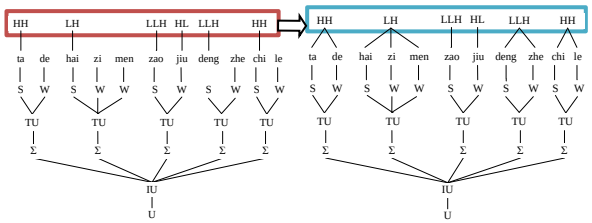

Figure 6: *Tonal coarticulation of accented syllables and the neighboring unaccented ones.*

This type of tone spreading is a progressive assimilatory process within a tonal unit. The tonal contour of TUs is the underlying phonological representation of the utterance prosodic pattern.

3.2. Intonation of mandarin speech prosody

In the following, we present the weighting and the grouping functions of intonation and their respective prosodic cues.

3.2.1. The weighting function

The weighting function of intonation marks the highlighted status of specific information from the utterance background. It was mentioned above that the prominent syllables were judged by 15 native subjects. We will now examine the features of accented syllables in comparison with those of the unaccented ones, by considering two acoustic parameters: syllable duration and aligned tonal feature.

A computation is firstly done on the correlation between accented syllables and their durations. The duration of each syllable is normalized according to the formula (1):

$$nT^{i}_{syllable} = \frac{T^{i}_{syllable}}{T_{mean}}, \quad i = 1, \cdots, m \qquad \textbf{(1)}$$

Where $T^{i}_{syllable}$ represents the i^{th} syllable duration in each utterance, m is the number of syllables in the utterance, and T_{mean} represents the average duration of the utterance. In this study, the 596 syllables (192 accented and 404 unaccented) in 60 utterances were normalized. The Pearson correlation reveals that the coefficient r between stressed and unstressed is 0.487** ($p \leq 0.01$). This confirmes that duration serves as an important cue in signalling the accented status of syllables in an utterance.

As observed in section 3.1, tone sandhi is an accent-related phenomenon, whereby the domain of the tone sandhi process is conditioned by prosodic hierarchy. Here we further evaluate this by analyzing the 58 sandhi-tone syllables found in the 60 test utterances, looking for possible correlation between the accentual status of syllables and the occurrence of tone sandhi. The Pearson correlation coefficient shows that r is -0.129** ($p \leq 0.01$). Although the number of sandhi-tone syllables concerns a small portion, i.e. only 9.7% of the total number of syllables in the data, the correlation between accented syllables and the tone-sandhi process confirms the tonal stability of strong syllables, as contrasted with the variable features of weak syllables.

3.2.2. The grouping function

The other fundamental linguistic function of intonation indicates prosodic boundaries. The boundary markers signal the finality or continuity of an utterance in connected speech.

The final syllable of an IU is often lengthened by the speakers to mark finality [20]. A correlation test between syllable position (sentence-final) and syllable duration in our study yielded a Pearson correlation coefficient of r = 0.311** ($p \leq 0.01$). This shows that the lengthened duration is closely related to the final position of the IU.

Pitch reset between successive IUs is also a reliable cue to define prosodic junctures, as speakers often drop the pitch level towards the end of an IU, with a final low tone indicating sentence finality. By contrast, when speakers start a new IU, they often initiate with a higher tone, marking the start of a new topic. In this study, the declination phenomenon in Mandarin utterances was investigated by comparing sentence final vs. non-final syllables, with z-score normalized Fø. As it happens, due to the down-drifting tendency of the global pitch level in each IU, the final syllable should have a lower Fø value than non-final syllables. The Pearson correlation test between the sentence-final syllable and the normalized Fø value shows an r coefficient of -0.238** ($p \leq 0.01$), which indicates that utterance-final syllables are more likely to have lower average pitch values than non-final syllables.

For the underlying annotation of the intonation [±terminal] tones, the H/L targets were employed, with the first tone of an IU represented as [H or [L (high vs. low initial tone, respectively), and the final boundary tone represented as H] or L] (high vs. low final tone). Since most of the sample utterances employed in the study are neutral productions with no strong emotion or emphasis, the melodic form of the utterance prosody mostly stems from the lexical tonal contours,

with the intonational boundary tones for prosodic phrasing superimposed at the initial and final part of the utterance.

The sample utterance (1) is a statement, so the terminal tone of the IU is annotated as L, while the initial tone is annotated as H. The pitch contours of tonal units together with the intonation boundary tones, compose the formal representation of the Mandarin IU. The prosodic form at the underlying phonological level of an utterance stems from the analysis of the prosodic function, with the procedure shown in the following figure:

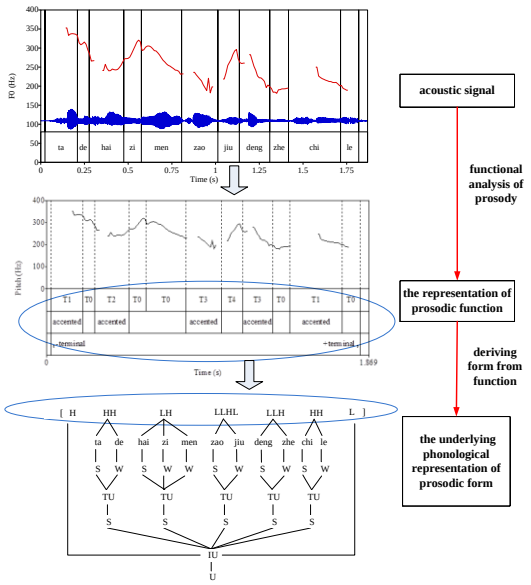

Figure 7: *Deriving the underlying phonological representation of prosodic form based on the functional analysis.*

The proposed representation fed into ProZed for evaluation. The annotation at the functional level provides the basic framework for the underlying phonological representation of the prosodic form. The lexical contours aligned on each tonal unit, together with the intonation boundary tones, are represented at the surface level with the INTSINT alphabet. These are the predicted manual annotations, to be tested by means of the Prozed tool to derive the synthesized output.

4. Applying ProZed to Mandarin speech prosody

The implementation of ProZed requires prosodic annotation via two tiers in the Praat TextGrid: TU and IU. In ProZed implementation, TU and IU tiers fined the unit for encoding, respectively, the short-term pitch control, and the long-term pitch variation [3].

TU was discussed in section 3 as the unit conditioned by prosodic grouping. It is the domain where the tone sandhi process operates in Mandarin prosody. The pitch movement of the surface melodic contour contributed by lexical tones in each TU, as well as the intonation boundary tones of the overall IU, are annotated by the INTSINT symbols in the following figure:

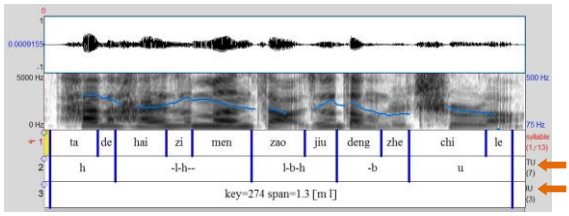

Figure 8: *Annotation of a mandarin utterance on TU tier and IU tier*

The symbolic representation is annotated at both the TU and the IU tier. The coding of the utterance prosody is the surface representation of prosodic form, as derived via the intermediate levels from the functional analysis of prosody.

The "*[m*" and "*l]*" symbols at the IU tier respectively code the utterance initial and final boundary tones. Since speakers generally start from their mid-level pitch range when producing a neutral utterance, thus "*[m*" symbol indicates the default initial tone of the utterance prosody. The final boundary tone of the utterance is defined according to the actual pitch tendency of the intonation unit. A low-tone target "*l]*" marks the final boundary, as based on the declining contour of the overall pitch pattern.

In the annotation window, besides annotation of the INTSINT tonal segments, two further parameters ("key" and "span") are specified at the IU interval tier. The key parameter corresponds to the mid pitch level in the speaker's current pitch range, while the span parameter indicates the pitch range between the maximum and minimum values symmetrically above and below the speaker's key level. The acoustic values of the key parameter (in Hertz) and the pitch span (in octavess) of each IU can be automatically derived by means of the MOMEL algorithm.

The surface coding of the utterance melodic contour specified at the TU and IU tiers are fed into the ProZed implementation for speech synthesis [3, 4]. In the following Figure 6, the green-colour points are the resulting pitch targets, which are interpolated by a quadratic spline curve, leading to a smooth and continuous pitch contour as the synthesized output:

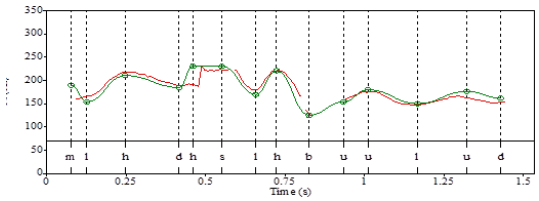

Figure 9: *Synthesized output of the utterance prosody*

One can easily check that the synthesized pitch contour (green-colour) closely follows the pitch movement of the original utterance (discontinuous red curve).

The present study employed acoustic data, symbolic coding, and speech synthesis to obtain a mapping between prosodic function and prosodic form. An analysis-by-synthesis study of Mandarin speech prosody was thus carried out. The symbolic representation of the speech melody helps reduce the observable complexity of a large quantity of data to a more simplified model. Through cyclic multi-level derivation, the synthesized data not only captures the pitch features of the original utterance at the physical level, but can also be interpreted with respect to the prosodic functions. It can thus provide useful insights for the annotation and modelling of the Mandarin speech prosody.

5. Acknowledgement

This work was supported by the first author's PhD scholarship from Scuola Normale Superiore di Pisa, Italy, and also by the National Basic Research Program (973 Program) of China (No. 2013CB329301), CASS innovation project 'Key Laboratory of Phonetics and Speech Science'.

6. References

[1] D.J. Hirst, and A. Di Cristo, *Intonation Systems. A Survey of Twenty Languages*. Cambridge: Cambridge University Press, 1998.

[2] D.J. Hirst, and C. Auran, "Analysis by synthesis of speech prosody: the ProZed environment," *Proceedings of the 9th Interspeech Conference*, Lisbon, pp. 3225-3228, 2005.

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

[4] D.J. Hirst, "ProZed: a speech prosody analysis-by-synthesis tool for linguists," *Proceedings of the 17th International Congress of Phonetic Sciences*, Hongkong, pp. 15-18, 2012.

[5] D.J. Hirst, "ProZed: a speech prosody editor for linguists, using analysis-by-synthesis," In K. Hirose & J. H. Tao (eds). *Speech Prosody in Speech Synthesis - Realizing High Quality and Flexible Control in Prosody for Speech Synthesis*. Springer Verlag; Berlin, Heidelberg, pp. 3-17, 2014.

[6] D.J. Hirst, "Form and function in the representation of speech prosody," In K. Hirose, D.J. Hirst, Y. Sagisaka (eds.), *Quantitative Prosody Modelling for Natural Speech Description and Generation* (=Speech Communication 46 (3-4)), pp. 334-347, 2005.

[7] D.J. Hirst, A. Di Cristo, and R. Espesser, "Levels of representation and levels of analysis for the description of intonational systems," In M. Horne (ed.), *Prosody: Theory and Experiment*. Dordrecht: Kluwer Academic Press, pp. 51-88, 2000.

[8] D.J. Hirst, and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic spline function," *Travaux de l'Institut de Phontique d'Aix*, 15: pp. 71-85, 1993.

[9] D.J. Hirst, "A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation," *Proceedings of the 16th International Congress on Phonetic Sciences*, Saarbrucken, pp. 1233-1236, 2007.

[10] N. Zhi, D.J. Hirst, and P.M. Bertinetto, "Automatic analysis of the intonation of a tone language. Applying the momel algorithm to spontaneous standard Chinese," *Proceedings of the 11th Interspeech Conference*, Makuhari, Japan, 2010.

[11] P. Boersma, and D. Weenink, Praat: a system for doing phonetics by computer. Available at http://www.praat.org, 1992-2016.

[12] S.H. Peng, M.K. Chan, C.Y. Tseng, T. Huang, O.J. Lee, and M.E. Beckman, "Towards a pan-Mandarin system for prosodic transcription," In S.A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. New York: Oxford University Press, pp. 230-270, 2005.

[13] E. Gårding, "Intonation in Swedish," *Working Papers 35*, Department of Linguistics, Lund University, pp. 63-88, 1989.

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

[15] M. Yip, *Tone*. Cambridge: Cambridge University Press, 2002.

[16] M.Y. Chen, *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge: Cambridge University Press, 2004.

[17] P.M. Bertinetto, C. Bertini, and N. Zhi, "A comparision of accentual features between Chinese and Italian speech," *Proc.6th International Conference of Speech Prosody*, Shanghai, China, pp. 520-523, 2012.

[18] N. Zhi, *A Study on Form and Function of Prosody based on Acoustics, Interpretation and Modelling*. Beijing World Publishing Corporation, Beijing, 2015.

[19] E.O. Selkirk, Phonology and Syntax: The Relation between Sound and Structure. Cambridge: MIT Press, 1984

[20] X.N. Shen, *The Prosody of Mandarin Chinese*. Berkeley: University of California Press, 1990.