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

Prosody Modeling of Spontaneous Mandarin Speech and Its Application to Automatic Speech Recognition

Cheng-Hsien Lin [1, 3], Meng-Chian Wu[1], Chung-Long You[1], Chen-Yu Chiang[2], Yih-Ru Wang[1], Sin-Horng Chen[1]

[1]Dept. of Electrical and Computer Engineering, National Chiao Tung University, Taiwan
[2]Dept. of Communication Engineering, National Taipei University, Taiwan
[3]Information & Communications Research Labs, Industrial Technology Research Institute, Taiwan

{jslin.cm98g, jameslibra.cm02g, yclong.cm02g}@nctu.edu.tw, cychiang@mail.ntpu.edu, yrwang@cc.nctu.edu.tw, schen@mail.nctu.edu.tw

Abstract

A prosody-assisted ASR approach for spontaneous Mandarin speech is proposed. It employs the joint prosody labeling and modeling algorithm proposed previously to construct a hierarchical prosodic model (HPM) and uses it in two-stage speech recognition. A word lattice is first generated by the HMM method using tri-phone AM and bigram LM. Then, the lattice is extended by replacing LM to a trigram model. A rescoring process is applied in the second stage to sequentially add factor POS and PM LMs, and the HPM. The method is evaluated on the MCDC database comprising 8 dialogues of 16 speakers with length of 9.09 hours. Error rates of syllable/character/word were reduced from 35.6/40.2/45.1% by the baseline trigram HMM method to 32.4/36.5/41.8% by the proposed method. The improvement is reasonably good as considering the WER upper-bound of 13.4% for the word lattice owing to the high OOV rate of the database. By error analysis, we find that many tone recognition errors and word segmentation errors were corrected. Besides, some information of the testing utterance was also obtained by the ASR, including POS of word, PM, tone of syllable, break type of syllable juncture, and prosodic state of syllable.

Index Terms: prosody-assisted ASR, prosody modeling, spontaneous Mandarin speech

1. Introduction

The use of prosodic information to assist in automatic speech recognition (ASR) is an interesting research topic. Prosody is referred to as the suprasegmental features of continuous speech, such as accentuation, prominence, tone and break, intonation, and rhythm. Prosodic information is known to be conveyed in the prosodic-acoustic features including pitch, energy, duration, and pause of spoken utterances. Prosody is also known to closely correlate with the linguistic features. The general approach of using prosodic information in ASR is to firstly identify important prosodic cues, that are closely correlated with linguistic features, such as accent, punctuation-related major break, and pre-boundary syllable lengthening. It then builds prosodic models from a large corpus to describe the relationships of those prosodic cues, linguistic features of text, and prosodic-acoustic features of speech utterance. Lastly, it incorporates the trained models into the ASR framework.

In the past, we have proposed a prosody-assisted ASR method [6-8] for Mandarin read speech. A hierarchical prosodic model (HPM) was built by a prosody labeling and modeling (PLM) algorithm from a prosody-unlabeled training corpus [1,6-8]. With the help of the HPM, the performance of the HMM-based Mandarin read-speech ASR were significantly improved. In this study, we want to extend the previous work to the ASR for Mandarin spontaneous speech. A preliminary study has been conducted to construct an HPM for Mandarin spontaneous speech by a modified PLM algorithm [9]. The modify PLM algorithm considers both the normal speech part and the disfluency speech part. The well-trained HPM will be used in this study to assist in spontaneous-speech ASR.

The remainder of the paper is stated as follows. Section 2 briefly introduces the spontaneous-speech HPM. Section 3 presents the proposed ASR framework. Experimental results are discussed in Section 4. Some conclusions are given in the last section.

2. The Hierarchical Prosodic Model

The preliminarily-trained hierarchical prosodic model (HPM) [9] is composed of 8 sub-models to describe various relationships among the prosodic-acoustic features of an utterance, the prosodic tags representing a four-layer prosodic structure of the utterance, and the linguistic features of the associated text. The basic modeling units used in this study are syllable (SYL) for normal speech parts and particular unit (PU) for disfluency part. PUs mainly comprise discourse marker (DM), particle, uncertain pronunciation, and over-lengthening syllable due to hesitation. Fig. 1 displays the four-layer prosodic structure used in the HPM. It is a modified version of the structure utilized in read-speech prosody modeling [1,6,7,10]. It consists of four layers: syllable/particular unit (SYL/PU), prosodic word (PW), prosodic phrase (PPh), and breath group/prosodic phrase group (BG/PG). Two types of prosodic tags, break type of syllable juncture and prosodic state of syllable, are employed to represent the prosodic structure. A set of seven break types, **B**={*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4}, is used to delimit prosodic constituents of these four layers. Here, *B*0 and *B*1 represent an intra-PW boundary with adjacent syllables/Pus being tightly and normally coupled,

10.21437/SpeechProsody.2016-212

respectively; *B*2-1, *B*2-2 and *B*2-3 are PW boundaries with obvious F0 reset, perceived short pause and pre-boundary lengthening; and *B*3 and *B*4 represent major breaks with medium and long pause durations, respectively. Primitive prosodic-acoustic features used to specify the break type B_n are pause duration pd_n, energy-dip level ed_n, normalized pitch jump pj_n, and normalized duration lengthening factor dl_n of syllable/PU juncture *n*.

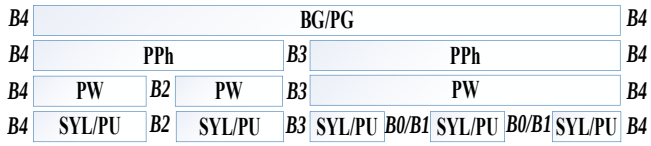

Figure1: *The 4-layer prosodic structure*

The prosodic state tag is employed to characterize the prosodic-acoustic feature patterns of prosodic constituents. Three types of prosodic states, $\mathbf{P}=\{p_n,q_n,r_n\}$, are used respectively for the three prosodic-acoustic features of syllable logF0 contour $\mathbf{sp}_n$, syllable duration sd_n, and syllable energy level se_n.The HPM describes the relationships of the prosodic tag sequence $\mathrm{T}=\{B_n,p_n,q_n,r_n\,|\,n=1\cdots N\}$, the prosodic-acoustic feature *sequence* $\mathbf{A}=\{sp_n,sd_n,se_n,pd_n,ed_n,pj_n,dl_n|n=1\cdots N\}$, and the linguistic feature sequence $\mathbf{L}=\{l_n,t_n,s_n,f_n,pos_n,cl_n|n=1\cdots N\}$, and is expressed by

$$\begin{aligned}&P(\mathbf{T}|\mathbf{A},\mathbf{L},\Lambda)\\&\approx\prod_{n=1}^{N}\begin{pmatrix}p(\mathbf{sp}_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1},pos_n)\,p(sd_n|B_{n-1}^{n},q_n,t_n,s_n,pos_n,cl_n)\\p(se_n|B_{n-1}^{n},r_n,t_n,f_n,pos_n)\end{pmatrix}\\&P(p_1)P(q_1)P(r_1)\prod_{n=2}^{N}P(p_n|p_{n-1},B_{n-1})P(q_n|q_{n-1},B_{n-1})P(r_n|r_{n-1},B_{n-1})\\&\prod_{n=1}^{N-1}\left(P(pd_n,ed_n,pj_n,dl_n|B_n,\mathbf{l}_n)P(B_n|\mathbf{l}_n)\right)\end{aligned}\tag{1}$$

where $\mathbf{l}_n$, t_n, s_n, f_n, pos_n, and cl_n denote, respectively, reduced linguistic feature, tone, base-syllable, final, part-of-speech, and contraction/lengthening tag of syllable/juncture *n*; $p(\mathbf{sp}_n\,|\,B_{n-1}^{n},p_n,t_{n-1}^{n+1},pos_n)$, $p(sd_n\,|\,B_{n-1}^{n},q_n,t_n,s_n,pos_n,cl_n)$, and $p(se_n\,|\,B_{n-1}^{n},r_n,t_n,f_n,pos_n)$ are, respectively, the syllable pitch contour, duration, and energy-level sub-models; $P(p_n|p_{n-1},B_{n-1})$, $P(q_n|q_{n-1},B_{n-1})$ and $P(r_n|r_{n-1},B_{n-1})$ are the pitch, duration and energy prosodic-state sub-models; $P(pd_n,ed_n,pj_n,dl_n\,|\,B_n,\mathbf{l}_n)$ is the break-acoustics sub-model; $P(B_n|\mathbf{l}_n)$ is the break-syntax sub-model; and Λ denotes the set of model parameters. These 8 sub-models have the same designs as their counterparts in the HPM for read-speech prosody modeling except that separate models are used for SYL and PU in the three syllable-based prosodic-acoustic sub-models. Specifically, the following sub-models are changed:

$$\begin{aligned}&p(\mathbf{sp}_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1},pos_n)\\&=\begin{cases}N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}^{f}_{B_{n-1},tp_{n-1}}+\boldsymbol{\beta}^{b}_{B_{n-1},tp_{n-1}}+\boldsymbol{\beta}_{pos_n}+\boldsymbol{\beta}_{B_{n-1},B_n}+\boldsymbol{\beta}_{p_n}+\boldsymbol{\mu}_{\mathbf{sp}},\mathbf{R}^{r}_{\mathbf{sp}}), & \text{for SYL}\\N(\mathbf{sp}_n;\boldsymbol{\beta}_{pr_n}+\boldsymbol{\beta}_{B_{n-1},B_n}+\boldsymbol{\beta}_{pr\_p_n}+\boldsymbol{\mu}_{\mathbf{pr\_sp}},\mathbf{R}^{r}_{\mathbf{pr\_sp}}), & \text{for PU}\end{cases}\end{aligned}\tag{2}$$

$$\begin{aligned}&p(sd_n|B_{n-1}^{n},q_n,t_n,s_n,pos_n,cl_n)\\&=\begin{cases}N(sd_n;\gamma_{cl_n}+\gamma_{t_n}+\gamma_{s_n}+\gamma_{pos_n}+\gamma_{B_{n-1},B_n}+\gamma_{q_n}+\mu_{sd},R^{r}_{sd}), & \text{for SYL}\\N(sd_n;\gamma_{cl_n}+\gamma_{pu_n}+\gamma_{B_{n-1},B_n}+\gamma_{pu\_q_n}+\mu_{pu\_sd},R^{r}_{pu\_sd}), & \text{for PU}\end{cases}\end{aligned}\tag{3}$$

$$\begin{aligned}&p(se_n|B_{n-1}^{n},r_n,t_n,f_n,pos_n)\\&=\begin{cases}N(se_n;\alpha_{t_n}+\alpha_{f_n}+\alpha_{pos_n}+\alpha_{B_{n-1},B_n}+\alpha_{r_n}+\mu_{se},R^{r}_{se}), & \text{for SYL}\\N(se_n;\alpha_{pr_n}+\alpha_{pr\_r_n}+\alpha_{B_{n-1},B_n}+\mu_{pr\_se},R^{r}_{pr\_se}), & \text{for PU}\end{cases}\end{aligned}\tag{4}$$

where β_x, γ_x and α_x represent affecting patterns (APs) of affecting factor *x* for syllable/PU pitch contour, duration and energy models, respectively; $\boldsymbol{\mu}_x(\mu_x)$ and $\mathbf{R}_x(R_x)$ denote respectively the global mean vector (mean) and the covariance matrix (variance) of modeling residual. It is worthy to note that common break type and prosodic state are used for both the normal speech parts and the disfluency parts. Accordingly, common sub-models related to prosodic tags of break type and prosodic states for both normal speech part and disfluent speech part are used. This can make the upper-layer prosody modeling cover the whole utterance without being interrupted by the normal-disfluency change of speech flow.

The PLM algorithm [1,9] is employed to simultaneously train the HPM and label the corpus with prosodic tags. It is a sequential optimization procedure to iteratively update model parameters and re-label prosodic tags based on the following criterion

$$\mathbf{T}^{*},\Lambda^{*}=\arg\max_{\mathbf{T},\Lambda}P(\mathbf{T}|\mathbf{A},\mathbf{L},\Lambda)\tag{5}$$

A heuristic initialization process built based on the general linguistic knowledge about spontaneous Mandarin speech is suggested to make the prosodic model training fully automatic from an unlabeled speech database.

3. The proposed ASR Framework

Basically, the task of the proposed ASR decoding is to find the best linguistic transcriptions $\Lambda_l=\{\mathbf{W},\mathbf{POS},\mathbf{PM},\mathbf{PU}\}$, prosody tags $\mathbf{T}=\{\mathbf{B},\mathbf{P}\}$, and acoustic segmentation Υ_s for the given input acoustic features $\mathbf{X}_a=\{\mathbf{X}_s,\mathbf{A}\}$ based on the following MAP criterion:

$$\begin{aligned}\Lambda_l^{*},\mathrm{T}^{*},\Upsilon_s^{*}&=\arg\max_{\Lambda_l,\Lambda_p,\Upsilon_s}P(\mathbf{W},\mathbf{POS},\mathbf{PM},\mathbf{B},\mathbf{P},\Upsilon_s|\mathbf{X}_s,\mathbf{A})\\&=\arg\max_{\Lambda_l,\mathrm{T},\Upsilon_s}P(\mathbf{W},\mathbf{POS},\mathbf{PM},\mathbf{B},\mathbf{P},\Upsilon_s,\mathbf{X}_s,\mathbf{A})\end{aligned}\tag{6}$$

where $\mathbf{W}=\{w_1^M\}$ is a word sequence; $\mathbf{POS}=\{pos_1^M\}$ is a POS sequence; $\mathbf{PM}=\{pm_1^M\}$ is a PM sequence; *M* is the total number of words; $\mathbf{B}=\{B_1^N\}$ is thebreak type sequence ; $\mathbf{P}=\{\mathbf{p},\mathbf{q},\mathbf{r}\}$ is the prosodic state sequence with $\mathbf{p}=\{p_1^N\}$, $\mathbf{q}=\{q_1^N\}$ and $\mathbf{r}=\{r_1^N\}$; *N* is the total number of syllables ; $\mathbf{X}_s$ is a frame-based spectral feature sequence (i.e., MFCCs

and their derivatives); and $\mathbf{A}=\{\mathbf{X},\mathbf{Y},\mathbf{Z}\}$ is a prosodic-acoustic feature sequence with **X**, **Y**, and **Z** representing sequences of syllable-related features, inter-syllable-related features, and inter-syllable differential features, respectively.

In this study, a two-stage recognition approach shown in Fig.2 is adopted. In the first stage, a word lattice is generated by the conventional HMM method with a triphone-based AM and a word-bigram LM. In the second stage, a rescoring scheme basing on Eq.(6) is used to find the best recognized word sequence.

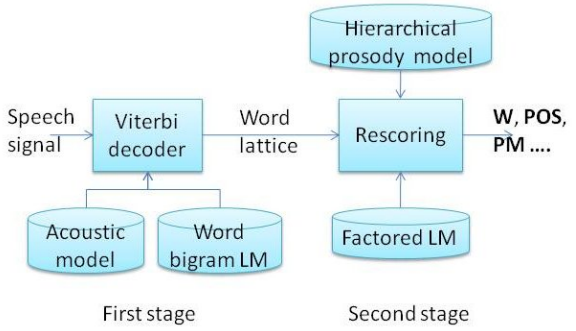

Figure 2: *A block diagram of the two-stage ASR decoding.*

In the rescoring process, the scores of many models will be combined. A log-linear score combining scheme shown below is used:

$$L(S,\Lambda_\alpha)=\log C(\Lambda_\alpha)+\sum_{j=1}^{K}\alpha_j \log p_j \qquad (7)$$

where $S=[p_1\cdots p_K]$ is a score vector formed from model score p_j ; $\Lambda_\alpha=[\alpha_1\cdots\alpha_K]$ is a weighting vector; and $C(\Lambda_\alpha)$ is a normalization factor. The discriminative model combination algorithm [2] is adopted to find the optimal weighting vector using the development set.

4. Experiments

The proposed ASR method is tested on MCDC corpus [3] containing eight dialogues collected by the Institute of Linguistics of Academia Sinica, Taiwan. Its total length is about eight hours. The eight dialogues were uttered by nine female and seven male speakers and transcribed into Chinese texts with some other tags including discourse marker (DM), particles, and pauses by professional linguist annotators.

Considering about the data sparseness problem, a set of triphone HMM model was firstly trained from TCC300 corpus [4] and then adapted using 90% MCDC data by the well-known Hidden Markov Toolkit (HTK) [5]. The acoustic feature vector is composed of 12 MFCCs and their delta and delta-delta terms, 1delta energy and 1 delta-delta energy.

In the acoustic model, several context independent HMM models are built for modeling PUs and paralinguistic phenomenon to deal with the disfluency in spontaneous speech. For PU part, 6 particle models (HO、EI、HAN、HEN、HEIN and MHM) plus 2 filler models (one for English, the other for Japanese and uncertain speech sound) are independently trained from MCDC data while other particles and markers are merged with normal syllables according to their Chinese characters. For paralinguistic part, 6 models are also trained for BREATHE、CLEAR_THROAT、LAUGH、NOISE、SMACK and SWALLOW from MCDC, and they are set to be optional units following every word in the recognition network.

A text corpus was employed to train both the word-bigram LM and the factored LM to be used respectively in the first- and second-stage decoding. The corpus contains in total about 440 million words and is formed by combining the following three corpora: (1) Sinorama: a news magazine with 9.87 million words; (2) NTCIR: an IR test bench consisting of several domains with 124.4 million words; (3) Sinica Corpus: general text corpus collected for language analysis with 4.8 million words; (4) Chinese Gigaword corpus: international news corpus with 262.5million words from LDC; and (5) Wikipedia page with 82.9 million words. A CRF-based tagger is employed to segment the corpus into word/POS sequences. For simplicity, PMs are categorized into four classes: comma, period, dot, and non-PM. A 60,005-word lexicon including all particles and markers was constructed based on the word frequency. LMs are also adapted using 90% MCDC corpus. It is noted that all paralinguistic marks are excluded in LM training and particles are replaced with similar pronounced characters beforehand.

In the first-stage decoding, a word lattice was produced using the triphone HMM and word bigram LM, and52.96% word accuracy (WAcc) was achieved; 54.92% WAcc was then obtained when a trigram LM is applied to expand and rescore the lattice. The word coverage rate of lattice is 86.63%, and this is the oracle performance of the second-stage decoding. In the second-stage rescoring processing, we first evaluated the performance of rescoring using factored LM (FLM) without involving any prosodic information. The recognition performance is shown in Table 1. The error rates of base-syllable, character, and word were 35.6%, 40.2%, and 45.1%. By incorporating the prosodic model, the performance improved to32.4%, 36.5%, and 41.8%. We also found the error rates of POS and PM were 10.1% and 13.7% for the proposed ASR method. It is worthy to mention that the tone recognition error is reduced from 4.5% to 3.8% (15.6% relative).

Table 1. *Error rates of base-syllable, character, word, POS, PM and TONE recognition attained on the MCDC corpus.*

	SER	CER	WER	POS	PM	TONE
Baseline (factored LM)	35.6	40.2	45.1	10.2	11.5	4.5
Baseline + prosody model	32.4	36.5	41.8	10.1	13.7	3.8

By an error analysis, we found that the performance improvement mainly resulted from the corrections of both tone recognition errors and word segmentation errors. Fig.3 shows an example of error correction. It can be found that word segments of wo3 and liu4shi2 (“me” and “sixty” in English) in the baseline result were corrected to segments wu4liu2 and shi4 (“logistics” and “is” in English), and the tone pattern 3-4-2 was also corrected to 4-2-4. It appears that the HPM can help to distinguish the correct word boundaries when incorporating

these intra/inter syllable prosodic and linguistic features in rescoring.

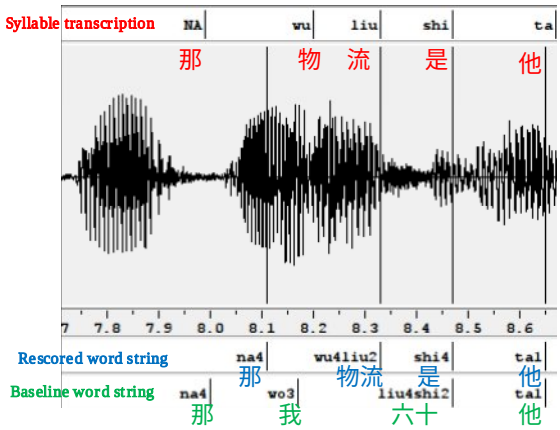

Figure 3: *An example of word segmentation and tone correction*

Moreover, the HPM was also found to be helpful for recognizing disfluency event in speech. Table 2 shows the example of decoded string. In this example, the repetition event is indicated by three tags: start (^), interrupt point (*), and end (#). The misrecognized repetition words "dui a" "dui a" were recovered after prosodic rescoring. These experiment results suggest that the prosody information which HPM introduced could be beneficial for spontaneous-speech ASR.

Table 2. *The example of decoded string for disfluent speech*.

Reference	^(對阿)*(對阿)# 阿 像 這 ^(dui a)*(dui a)# a xiang zhe
Baseline	^(NULL)*(NULL)# 那 了 好像 在 這 ^(NULL)*(NULL)#na le hao xiang zai zhe
Rescored	^(對阿)*(對阿)# 好像 在 這 ^(dui a)*(dui a)#hao xiang zai zhe

5. Conclusions

In this study, the prosody-assisted read-speech ASR approach proposed previously has been successfully extended to the ASR for Mandarin spontaneous speech. Comparing with the syllable/character/word error rates of 35.6/40.2/45.1% by the baseline trigram HMM-based method, the proposed approach achieved the improved performance of 32.4/36.5/41.8%. These improvements are reasonably good as considering the WER upper-bound of 13.4% confined by the word lattice generated in the first-stage recognition due to the high OOV rate of the database. By error analysis, we find that the improvement mainly resulted from the corrections on tone recognition errors and word segmentation errors. An additional advantage of the proposed approach lies in the exploration of some information of the testing utterance by the ASR, including POS of word, PM, tone of syllable, break type of syllable juncture, and prosodic state of syllable.

Further study to consider the compensation of the effect of high speaking rate of spontaneous speech on ASR is worth doing in the future. The introduction of the speaking rate-dependent HPM [11] to spontaneous-speech prosody modeling maybe a good starting point.

6. Acknowledgements

This work was mainly supported by Ministry of Science and Technology under Contract "MOST 101-2221-E-009-078-MY3" and partially by the Ministry of Economic Affairs of Taiwan. The authors want to thank Dr. S. C. Tseng of Academia Sinica for providing MCDC corpus.

7. References

[1] C. Y. Chiang, S. H. Chen, H. M. Yu, and Y. R. Wang, " UnsupervisedJoint Prosody Labeling and Modeling for Mandarin Speech " J. Acoust.Soc. Am., vol.125, no.2, pp.1164-1183, 2009.

[2] P. Beyerlein, "Discriminative model combination," in *Proc. ICASSP* 1998, pp. 481-484.

[3] S. C. Tseng, " Processing spoken mandarin corpora, " TraitementAutomatique des Langues, vol.45, no.2, pp.89 – 108, 2004.

[4] Mandarin microphone speech corpus – TCC300, http://www.aclclp.org.tw/use_mat.php#tcc300edu.

[5] "HTK Web-Site", http://htk.eng.cam.ac.uk.

[6] Jyh-Her Yang, Ming-Chieh Liu, Hao-Hsiang Chang, Chen-Yu Chiang, Yih-Ru Wang, and Sin-Horng Chen, "Enriching Mandarin speech recognition by incorporating a hierarchical prosody model," in Proc. ICASSP 2011, Plaque, Czech, May, 2011, pp 5052-5055.

[7] Sin-Horng Chen, Jyh-Her Yang, Chen-Yu Chiang, Ming-Chieh Liu and Yih-Ru Wang, "A New Prosody-Assisted Mandarin ASR System," IEEE Trans. on Audio, Speech and Language Processing, Vol. 20, No. 6, pp. 1669-1684, August 2012

[8] Chen-Yu Chiang, Sabato Marco Siniscalchi, Yih-Ru Wang, Sin-Horng Chen, Chin-Hui Lee, "A Study On Cross-Language Knowledge Integration In Mandarin LVCSR," in Proc. ISCSLP, Dec. 2012, Hong Kong

[9] Yu-Lun Chou, Chen-Yu Chiang, Yih-Ru Wang, Hsiu-Min Yu, Sin-Horng Chen,"Prosody Labeling and Modeling for MandarinSpontaneous Speech", submitted to ICASSP2016

[10] C.-Y. Tseng, S.-H. Pin, Y.-L. Lee, H.-M. Wang, and Y.-C. Chen, "Fluent speech prosody: Framework and modeling," Speech Communication, 46, pp. 284-309, 2005.

[11] Sin-Horng Chen, Chiao-Hua Hsieh, Chen-Yu Chiang, Hsi-Chun Hsiao, Yih-Ru Wang, Yuan-Fu Liao and Hsiu-Min Yu, "Modeling of Speaking Rate Influences on Mandarin Speech Prosody and Its Application to Speaking Rate-controlled TTS," IEEE/ACM Trans. on Audio, Speech and Language Processing, Vol.22, No.7, pp.1158-1171, July 2014.