HOW PROSODIC CUES COULD LEAD TO INFORMATION CENTER IN SPEECH - AN ALTERNATIVE TO ASR

Chao-yu Su [1, 2, 3] *& Chiu-yu Tseng* [1]

1 Institute of Linguistics, Academia Sinica, Taiwan
2 Taiwan International Graduate Program, Academia Sinica, Taiwan
3 Institute of Information Systems and Applications, National Tsing Hua University, Taiwan
cytling@sinica.edu.tw

ABSTRACT

It has been reported in ASR literature that prosody helps retrieve important textual information by word. We therefore believe that prosodic information in the speech signal could be used to facilitate speech processing more directly. The prosodic word, a perceptually identifiable unit which is usually slightly larger in size than lexical word, can be a possible alternative to help locate important information in speech. Acoustic analysis across labels of perceived prosodic highlighted part in prosodic words and semantic foci in words are compared. The results demonstrate that prosodic highlights occur before targeted key information and function as advanced prompts to outline upcoming sematic foci ahead of time. Semantic saliency of targeted words are thus enhanced beforehand while correct anticipation can be facilitated prior to detailed lexical processing. Further automatic identification approach of key content by prosodic features also shows the possibility to retrieve important information through prosodic words. We believe the results demonstrate that not all information is equally important in speech, locating information center is the key to speech communication, and the contribution of prosody is critical.

Index Terms—*prosodic word, acoustic analysis, semantic foci, prosodic highlight, speech understanding, bypass ASR*

1. INTRODUCTION

Speech prosody, generally refers to modulations of fundamental frequency, duration, and amplitude in the speech signal, is regarded as a major determinant of the form and meaning of spoken language in terms of comprehensibility [1, 2]. Therefore, prosody has been applied to spoken Language Understanding (SLU) tasks by many researches [3, 4, 5, 6]. However, most SLU tasks using prosodic information are based on ASR (automatic speech recognition) output with succeeding SLU and NLP fine-tuned/optimized individually. One recent study argued that, without implementing ASR or any deeper linguistic analysis which is text based, pitch accent, one of the major prosodic features, could retrieve where the most important information is located in text [7]. The study examined the correlation between pitch accents and words that are annotated with semantic labels (semantic slots) on the Airline Travel Information System (ATIS) corpus; the results showed prosody could directly point towards the most salient information in text. The study adopted lexical semantic units, namely word units in text, as analysis scale/prediction response. However, other studies have pointed out that prosody should not be limited to the choice of lexical items only [8, 9], but how these items are related semantically, syntactically, and rhythmically instead. These studies, dated back to the 1980's, suggest that prosodic association among lower-level prosodic elements to form higher-level constituent which reflects syntactic structure and discourse association, are hierarchical in structure; linear sequencing is not sufficient. As it turned out, layering of discourse context and quantitative contribution of multiple layers of discourse hierarchy have been reported later [10], and summaries of related studies adopting the latter perspective could also be found in a recent review paper [11]. In short, defining 'prosody in context' as involving multiple layers of linguistic contextual factors as well as their influence and interaction with phonological forms and phonetic manifestations has already been accounted for with quantitative justifications. The consensus of these studies is such that under the complex interaction involved in continuous speech, prosody should be examined by including higher level contextual information instead of canonical linguistic units and/or their linear associations only. Therefore, to investigate how important information may be retrieved by way of prosody, the present study proposes that the scope/context under investigation should be enlarged at least to prosodic units, thereby encompassing more speech information with respect to semantic focusing into account.

We will report analysis of the prosodic word (PW) which is regarded as the lowest-level constituent in the prosodic hierarchy [9]. PW adopted here is defined by a perception based multi-phrase discourse prosody hierarchy (HPG) that accounts for both respective and cumulative contributions (by layers) to output continuous speech prosody [12, 13]. Each level of perceived prosodic boundaries was labeled

978-1-5386-3333-5/17/$31.00©2017 IEEE

manually and controlled for intra- and inter-transcriber consistency [12]. The PW boundaries were found to be highly consistent among transcribers (90-93%); the duration patterns of PWs and between-PW boundaries are also found to be systematic and predictable [13, 14], thereby substantiating the status of PW in the prosodic hierarchy. Further annotation of information status by perceived prosodic highlights revealed that PWs are also information planning units (IPU), differing in levels of prominence. The distribution of PWs as IPU revealed two major categories keyword 'KEY' and projector 'PJR'; the difference of their acoustic patterns is significant. It was found that 'KEY' (mostly nouns), produced with higher F0 and longer duration, functions as prosodic indexing of the location of key information itself whereas 'PJR' (all other POS), produced with higher F0 only, functions as prosodic projecting and advance prompting of soon-to-arrive focal information [15, 16]. Following the above findings, the present study assumes that salient PWs perceived as PJR/KEY could both signal where important information is in PW scale.

The goal of the present study is to examine prosodic patterns by the two different annotations/scales, namely word and PW as well as compare their prosodic saliency with respect to semantic density. In other words, PJR/KEY at PW level plays a role of advanced prompts to facilitate anticipation of upcoming important content. Similarly, the semantic foci (SF)/key information in lexical unit words can be treated as a meta-unit embedded in information unit PJR/KEY, and PW labeled as PJR/KEY would facilitate correct access of embedded semantic foci. To prove the assumption, the present study will examine acoustic correlates F0, duration and intensity of PJR/KEY at PW level and semantic (information) labels at word level in continuous discourse using native speech of read English. Prosodic highlight indexes and semantic labels are annotated separately. The former is perceptual labeling of prosodic units PW from speech data; the latter is manual labeling of word units of the same data in text form. In the following analysis we will first show how much overlap exists between PJR&KEY by PW and SF by word. Then we will show results how acoustic patterns by F0, duration and intensity correspond to prosodic highlight indexes and semantic labels, respectively. In addition, the present study will also examine performance of identifying key content by PJR&KEY in PW as well as SFs in word, respectively through artificial neural network and discusses their relationship.

2. TEXT, SPEECH MATERIALS AND RESPECTIVE ANNOTATION

The materials of English speech analyzed are native speech of story reading “The North Wind and the Sun” (henceforth NW&S) and “The Cinderella Fairy Tale” (henceforth Cinder) from the AESOPILAS and AESOP2-ILAS [13, 14] By text terms, NW&S includes a total of 113 words (144 syllables) in 3 paragraphs (5 sentences with 8 independent clauses and 5 dependent clauses); Cinder includes a total 759 words (1,000 syllables) in 14 paragraphs (82 sentences with 93 independent clauses and 49 dependent clauses). Speech data of NW&S and Cinder are from 11 (5M/6F) and 10 (5M/5F) L1 North American English speakers, respectively. The text data was tagged by semantic foci; both sets of speech data were tagged in separate layers by perceived prosodic highlights.

2.1 Annotation for semantic foci on text

Semantic foci by syntactic and semantic specifications were narrow focus (NF), broad focus (BF) and non-focus (NonF) and manually annotated by a native English linguist to specify focus status.

2.2 Annotation for prosodic highlight indexes on speech

Perceived prosodic highlights are manually tagged by levels of prominence, discourse units by perceived boundary breaks, and information status by PJR/KEY. A total 3 layers of annotations are described in 2.2.1, 2.2.2 and 2.2.3.

2.2.1. Annotating discourse units by perceived boundaries and breaks-the 1st layer

Discourse units were manually tagged by 4 levels of perceived discourse prosodic boundaries B2 through B5; and 5 levels of between-boundary prosodic units are defined as the prosodic word (PW/B2), the prosodic phrase (PPh/B3), the breath group (BG/B4, a physio-linguistic unit constrained by change of breath while speaking continuously) and the multiple phrase speech paragraph (PG/B5). PW/B2 are used as prosodic units/boundaries in the present study for tagging important content in prosodic perspective.

2.2.2. Annotating prosodic highlight by levels of perceived prominence-the 2nd layer

The same speech data were manually annotated, in a separated layer, into a string of perception-based emphasis/non-emphasis tokens (ETs). The annotation for prominence is based on 3 relative degrees of perceived strength, following the definitions:

- E1 -- normal pitch, normal volume and clearly produced segments
- E2 -- raised pitch, louder volume and irrespective of the speaker’s tone of voice
- E3 -- higher raised pitch, louder volume and with the speaker’s change of tone of voice

By this annotation scheme, we emphasize the fact that the distinctions in prominences can be perceived consistently by only limited numbers of contrastive levels.

2.2.3 Annotating perceived prosodic highlight by information status-the 3rd layer

We categorize the ETs with actual emphases, namely those of E2 and E3, based on the corresponding information content of each token by PW (B2) as IPUs keyword 'KEY' and projector 'PJR'. The PWs falling out of the above two categories are categorized into 'Others'

2.2.4 Distribution of labels of prosodic highlight and semantic foci

By semantic labels and speaker, NW&S and Cinder yield a total of 1243 and 7590 labeled words, respectively. By prosodic highlight indexes, NW&S and Cinder yield a total of 590 and 1743 labeled PWs, respectively. For NW&S and Cinder, PWs are larger than lexical words; each PW contains 2.1 and 4.35 lexical words in average, respectively. Tables 1 and 2 summarize their further categorization by semantic and prosodic-information indexes.

Table 1. *A summary of semantic labels*

Semantic labels Speech paragraph	Non-F	BF	NF
NW&S	53.10%	43.36%	3.54%
Cinder	53.50%	41.22%	5.28%

Table 2. *A summary of perceived prosodic highlights*

Prosodic highlight indexes Speech paragraph	Others	PJR	KEY
NW&S	66.04%	13.58%	20.38%
Cinder	46.02%	18.92%	35.06%

3. METHOD

3.1. Overlapping rate

To examine the overlap between prosodic highlights (PJR/KEY) at PW level and semantic foci (BF/NF) at word level, BF/NF and PJR/KEY are defined as predictive condition and true condition, respectively. PJR/KEY containing more than 50% of BF/NF by number and BF/NF covered by PJR/KEY are defined as correct prediction and true, respectively. The between PJR/KEY-BF/NF precision, recall and overall overlap (accuracy) is derived.

3.2. **Classifiers for identifying key content by perceived prosodic highlight/semantic labels**

The multilayer perceptron (MLP), a feedforward artificial neural network model that maps sets of input data onto a set of prediction outputs [17, 18], is adopted as classifier to automatically identifying key content, PJR/KEY and BF/NF defined by respective prosodic and semantic labels. The MLP network is then trained to approximate ground truth label in training phase to predict learned/unlearned input set (inside/outside test). An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Each node is a neuron with a nonlinear activation function. The setup for MLP in the present study is listed in following: activation function - sigmoid, hidden layer size - 7, Epoch # - 200 and dropout - 0.3. 7/8 of dataset are randomly selected and used for training network and the others of dataset are used for outside testing set.

3.2.1. Acoustic features

F0, intensity and duration are used in the present acoustic analysis. The 3 features are z-normalized by discourse prosody units to remove speaker variation while duration is further normalized by segmental identities to remove duration difference due to intrinsic physical composition. The normalized acoustic features are segmented into words and PWs respectively; values of mean, maximum, minimum stand deviation and slope (only for F0 and intensity) are calculated for prosodic features by two different unit/scale. The edge context (preceding and following) by one word/PW unit is also calculated and included as parts of input features. As a result, each word/PW contains altogether 70 dimensions of prosodic features. Then the acoustic features extracted are aligned with prosodic highlight indexes in PWs and semantic labels in words respectively for acoustic analysis as well as identification task for prosodic/semantic important content.

4. RESULTS AND DISCUSSION

4.1. Overlap between perceived prosodic highlights and semantic foci

Table 3 presents recall, precision and overlap rate between prosodic highlights at PW level and semantic foci at word level by NW&S and Cinder. The results show that the overall overlap between prosodic highlights (PJR/KEY)-semantic foci (BF/NF) is about 82% in NW&S and 70% in Cinder.

Table 3. *Overlap between prosodic highlights and semantic foci*

Measure Speech paragraph	Recall	Precision	Overlap
NW	79.63%	84.31%	81.91%
Cinder	70.49%	70.08%	70.28%

4.1.1 Discussion

Although prosodic indexes and focus status are annotated separately, namely, at PWs in speech by perception/prosody and at words in text by semantic/syntactic, the results, however, show considerable overlap. In other words, the PWs highlighted in the speech signal often cover important information content in text, namely key words as expected. The results thus suggest that prosodic highlights PJR/KEY embedded in prosodic units are capable of outlining semantic foci in correlation with information structure.

4.2. Acoustic patterns by perceived prosodic highlight

Figure 1 and 2 show mean patterns of F0, intensity and duration by perceived prosodic highlight at PW scale by NW&S and Cinder, respectively. The red circles represent 'PJR/KEY' patterns which significantly differ from 'others'. The significant test of PJR/KEY in reference to 'others' is listed in Table 4, 5, 6.

For 'KEYs' in NW&S and Cinder, F0 and intensity are significantly different from 'others' by higher F0 and stronger intensity. Duration patterns at 'Keys' in Cinder show faster tempo which is not shown in NW&S. For 'PJRs', significant difference from ‘others’ is found by higher F0 and stronger intensity in NW&S as well as shorter duration in Cinder.

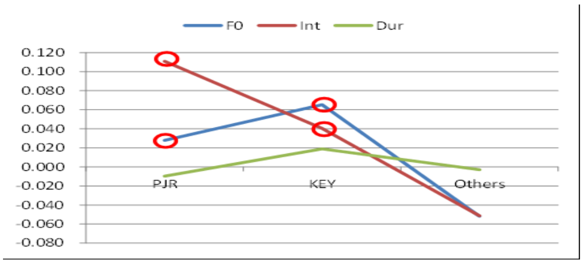

Figure 1: *Mean patterns of F0, intensity and duration by prosodic highlight indexes at PW scale by NW&S*

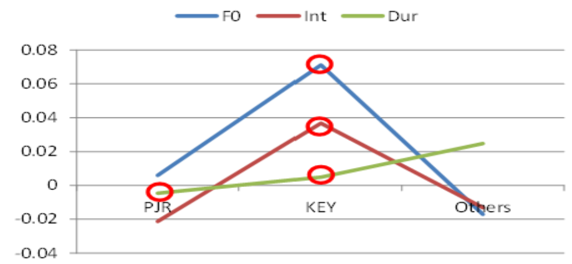

Figure 2: *Mean patterns of F0, intensity and duration by prosodic highlight indexes at PW scale by Cinder*

Table 4. *Two-sample t-test of PJR/KEY in reference to others by intensity by F0*

NW&S Cinder	Others	PJR	KEY
Others		H=1 (p<0.05)	H=1 (p<0.05)
PJR	H=0		
KEY	H=1 (p<0.05)		

Table 5. *Two-sample t-test of PJR/KEY in reference to others by intensity*

NW&S Cinder	Others	PJR	KEY
Others		H=1 (p<0.05)	H=1 (p<0.05)
PJR	H=0		
KEY	H=1 (p<0.05)		

Table 6. *Two-sample t-test of PJR/KEY in reference to others by duration*

NW&S Cinder	Others	PJR	KEY
Others		H=0	H=0
PJR	H=1 (p<0.05)		
KEY	H=1 (p<0.05)		

4.3. Acoustic patterns by semantic labels

Figure 3 and 4 show mean patterns of F0, intensity and duration by semantic labels specified by focus status at word scale by NW&S and Cinder, respectively. The red circles represent 'BF/NF' patterns which significantly differ from 'NonF'. The significant test of 'BF/NF' in reference to 'NonF' is listed in Table 7, 8, 9.

For both NW&S and Cinder, all prosodic features including F0, intensity and duration in words labeled as 'BF/NF' are significantly different from 'NonF'. The results show higher F0, stronger intensity and slower tempo are the major features of 'BF/NF' compared to 'NonF'.

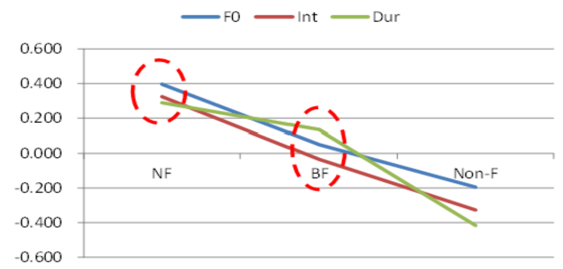

Figure 3: *Mean patterns of F0, intensity and duration by semantic labels at PW scale by NW&S*

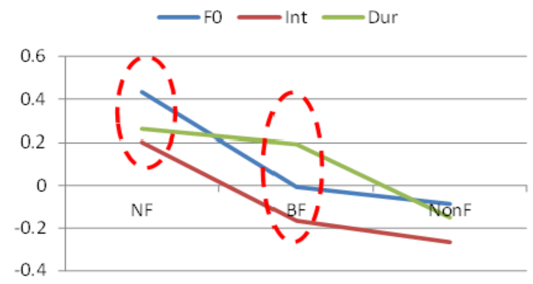

Figure 4: *Mean patterns of F0, intensity and duration by semantic labels at PW scale by Cinder*

Table 7. Two-sample t-test of BF/NF in reference to NonF by F0

NW&S Cinder	NonF	BF	NF
NonF		H=1 (p<0.05)	H=1 (p<0.05)
BF	H=1 (p<0.05)		
NF	H=1 (p<0.05)		

Table 8. *Two-sample t-test of BF/NF in reference to NonF by intensity*

NW&S Cinder	NonF	BF	NF
NonF		H=1 (p<0.05)	H=1 (p<0.05)
BF	H=1 (p<0.05)		
NF	H=1 (p<0.05)		

Table 9. *Two-sample t-test of BF/NF in reference to NonF by duration*

NW&S Cinder	NonF	BF	NF
NonF		H=1 (p<0.05)	H=1 (p<0.05)
BF	H=1 (p<0.05)		
NF	H=1 (p<0.05)		

4.3.1. Discussion

Compared to prosodic saliency of important content in perceived prosodic highlight (PJR/KEY) in 4.2, the key components by semantic labels (BF/NF) are more pronounced prosodically. However, note that 1) some PJRs/KEYs in PW level are significantly marked by F0 and intensity in 4.2, and 2) larger-size PJRs/KEYs often cover BFs/NFs in 4.1. We therefore assume that PJR/KEY at the larger-size PW level can be seen as a prosodic signal to prompt important content at smaller-size word level ahead of time. The prompting could facilitate advanced processing at higher-level PW to aid correct and detailed processing at lower-level word. With respect of the signal, prosodic boosting of PJR/KEY via larger-size unit could further enhance signal saliency of embedded BF/NF. To test the above assumption, the following section will further examine whether performance of automatic prediction for PJR/KEY is correlated to BF/NF.

.

4.4. Identifying key content by prosodic highlight indexes/semantic labels

Figure 5 shows performance of automatic identification of key content defined by prosodic highlights (PJR/KEY) and semantic foci (BF/NF), respectively, using acoustic features. The results show slightly better performance of identifying BF/NF in words than PJR/KEY in PWs. For inside test, the performance of BF/NF at words for NW&S and Cinder are 99% and 94%, respectively. On the other hand, the performance of PJR/KEY in PWs are 92.9% and 78.1%, respectively. For outside test, the performance of BF/NF in words are 86.4% and 78.5% for NW&S and Cinder, respectively, and the performance of PJR/KEY in PWs are 77.7% and 69% for NW&S and Cinder, respectively.

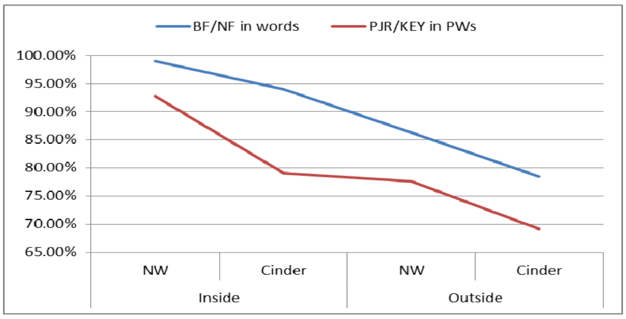

Figure 5: *Performance of automatic identifying key content using acoustic features for PJR/KEY in PWs and BF/NF in words respectively*

4.4.1. Discussion

The results of automatic identifying key content suggest that PJR/KEY (PW level) leads BF/NF (word level) in terms of prediction accuracy. The higher accuracy on PJR/KEY is positively correlated to accuracy increase of BF/NF. The results also show PJR/KEY performance achieves acceptable degree, namely, 70%~93%, even though it is not as good as BF/NF performance. The results further echo the assumption in 4.3.1 that PJR/KEY provides advance indications toward salient information location at larger-size PW, and that prosodic contrast/saliency of PJR/KEY boosts the embedded BF/NF ahead of time. We therefore believe that advance prompting of prosodic highlights in the speech signal thus helps facilitate accessibility to semantic foci in text.

5. GENERAL DISCUSSION

The above results suggest that in continuous speech prosodic highlights in PWs outline and enhance important semantic content in words which relates directly to information structure of speech content. The prosodic highlights are also found to play an antecedent role ahead of BF/NF in terms of the performance of identifying key content in continuous speech. These results thus account for how prompting/indexing via prosodic means, in this case highlighting, is a simple and straightforward strategy of speech planning available to help facilitate correct access of embedded semantic foci ahead of time. As suggested in previous studies [9,10], we also believe that in speech understanding tasks using prosodic units to locate information center in the speech signal and at the same time bypassing "force-aligned" linguistic units by ASR could be a less costly alternative to retrieve important content.

6. CONCLUSIONS

The present study conducted acoustic analysis using labels of prosodic highlights and semantic foci. Our novel approach to tackle information center in the speech signal directly reveals that units highlighted by prosody in fact function as advanced prompt to facilitate outlining semantic saliency ahead of time. This approach differs in spirit from the ASR approach that treats everything the speech signal with equal importance. We believe our results demonstrate that not all information is equally important in speech, locating information center is the key to speech communication, and the contribution of prosody is critical.

7. REFERENCES

[1] Cutler, A., Dahan, D., & van Donselaar, W. “Prosody in the Comprehension of Spoken Language: a literature review”. Language and Speech, 40, 141-201, 1997.

[2] Derwing, T. M. & Munro, M. J. “Accent, intelligibility, and comprehensibility: evidence from four L1s”. Studies in Second Language Acquisition, 20, 1-16, 1997.

[3] A. Batliner, B. M¨obius, G. M¨ohler, A. Schweitzer, and E. N¨oth, “Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground,” in Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark), vol. 4. ISCA, 2001, pp. 2285–2288, 2001.

[4] Veilleux, N. and Ostendorf ,M. “Prosody/parse scoring and its application in ATIS,” in In Proceedings of the ARPA Workshop on Human Language Technology, 1993, pp. 335–340, 1993.

[5] Shriberg, E., Stolcke, A., Hakkani-T¨ur,D. and Tur, G. “Prosodybased automatic segmentation of speech into sentences and topics,” Speech Communication, vol. 32, pp. 127–154, 2000.

[6] Shriberg, E., and Stolcke, A. “Prosody modeling for automatic speech recognition and understand Speech and Language Processing. Springer, 2004, pp. 105–114, 2004.

[7] Stehwien, S. and Vu, N-T., “Exploring the Correlation of Pitch Accents and Semantic Slots for Spoken Language Understanding”, Interspeech 2016, San Francisco US, 2016.

[8] Selkirk, EO. Phonology and syntax: the relation between sound and structure. Cambridge, MA: MIT Press; 1984.

[9] Nespor, M.; Vogel, I. Prosodic phonology. Dordrecht: Foris; 1986.

[10] Tseng, C-Y. “Beyond Sentence Prosody”. Interspeech. Makuhari, Japan, 2010.

[11] Cole, J. “Prosody in context: a review”, Language, Cognition and Neuroscience , Volume 30, 2015 - Issue 1-2: Prosody in Context, 2015.

[12] Tseng, C-Y. “The Prosodic Status of Breaks in Running Speech：Examination and Evaluation”. Speech Prosody 2002 667-670. Aix-en-Provence, France, 2002.

[13] Tseng, C-Y. Pin, S-H, Lee, Y-L, Wang, H-M and Chen Y-C. “Fluent speech prosody: framework and modeling. Speech Communication”, Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation 46(3-4): 284-309, 2005.

[14] Tseng, C-Y and Chang, C-H. 2007. “Pause or No Pause?—Phrase Boundaries Revisited”. The 9th National Conference on Man-Machine Speech Communication (NCMMSC 2007). China, 2007.

[15] Chen H K-Y, Fang, W-T, and Tseng C-Y. “Prosodic prompts and information planning units in continuous speech—Relative allocation and compensation of prosodic highlight”. The 12th Phonetic Conference of China (PCC 2016), (Jul. 25-26). Tongliao, China, 2016.

[16] Chen H K-Y, Fang, W-T, and Tseng C-Y. “Advance Prosodic Indexing - Acoustic realization of prompted information projection in continuous speeches and discourses”. ISCSLP 2016 - The 10th International Symposium on Chinese Spoken Language Processing, (Oct. 17-20). Tianjin, China, 2016.

[17] Haykin, S. “Neural Networks: A Comprehensive Foundation (2 ed.) ”. Prentice Hall. ISBN 0-13-273350-1, 1998.

[18] Collobert, R. and Bengio, S. “Links between Perceptrons, MLPs and SVMs”, Proc. Int'l Conf. on Machine Learning (ICML), 2004.