Bidirectional Text-To-Pronunciation Conversion with Word Stress Prediction for Ukrainian

Mykola Sazhok, Valentyna Robeiko

Speech Science and Technology Dept., IRTC, Kyiv, Ukraine
sazhok@gmail.com, valya.robeiko@gmail.com

Abstract

In this paper we consider the actual problem of text-to-pronunciation conversion that can be generalized to backward direction as well. The main purpose is to separate the operational code (instructions) and the data that is the way to obtain a flexible and convenient tool for the researcher. We propose a model describing regularities of relations between the orthographic spelling and phonetic symbols. Multi-decision symbol sequence correspondences carried out according to the model are equivalent to building a directed graph. Only about 30 generalized correspondence cover the literary pronunciation for Ukrainian. To convert text to pronunciation for spontaneous speech, we introduce additional model levels allowing the expert to build complex correspondences still working with relatively simple data structures. The other benefit of introduced levels is the possibility to convert numbers, symbols and abbreviations to their textual presentation within the same algorithm. Word stress is either pointed in accordance to the vocabulary or predicted automatically by the proposed text corpus-driven procedure. We also share experience of producing the phoneme sequences corresponding to different pronunciation ways and individual manners of the orthographic text.

Index Terms: grapheme-to-phoneme, word stress prediction, spontaneous speech, individual speech modeling

Introduction

Text-to-pronunciation and pronunciation-to-text are actual procedures for speech technology development. These kinds of conversions are obligatory in text-to-speech systems to form individualized phoneme transcriptions and in speech recognition systems to create pronunciation dictionaries and to organize advanced schemes of speech decoding [1]–[4].

To model text-to-pronunciation conversions (in both directions) we need to learn regularities between orthographic and phonetic symbols. Automatically driven regularities process a huge amount of transcribed words [5]. We rely on regularities extracted by an expert that is acceptable apparently for languages with pronunciation-based spelling, and Ukrainian is among them.

A decade ago, for Ukrainian, a text-to-pronunciation conversion procedure has been implemented in the program code that simulated pronunciation rules taken from the handbook [1]. To make the text-to-pronunciation or grapheme-to-phoneme (GTP) converter be a flexible tool for the researcher we needed to separate the program code and the data.

Historically, the opposite to GTP procedure has been implemented in a desired manner. Although we developed first versions of pronunciation-to-speech module several years ago as a part of our research [4], attempts to formalize the conversion were abandoned. Presented hereafter formal description for the model we accomplish with stress prediction procedure for Ukrainian that is driven from text corpus and a training material. Word stress position is not regular for Ukrainian and is crucial specifically for text-to-speech applications.

In next Section we describe the general model for multi-decision conversion between symbol sequences, in Section 2 we consider a word stress prediction algorithm, in Section 3 we describe the developed text-to-pronunciation system for Ukrainian and share the experience of work the system.

1.Model for the multi-decision conversion between symbol sequences

A key issue in modeling the conversion between symbol sequences is the question of how we define the correspondence between symbols of source and target sequences. We consider a finite sequence of source symbols $a_1^N=(a_1, a_2, \dots, a_n, \dots, a_N)$ where each element is taken from the alphabet of input symbols A. Let us construct the conversion of this sequence to a set of sequences for output symbols taken from the alphabet B.

Consider an elementary correspondence f that maps a subsequence of a_1^N, starting from its *n*-th symbol, to a symbol from the alphabet B or an empty symbol:

$$f\left(a_n^N\right)=b\ ,\ a_n^N\in \mathrm{Def}\left(f\right)\subset A,\qquad b\in B\cup\varnothing\ ,\ 1\le n\le N. \tag{1}$$

Note that (1) is applicable only for the specified source sequences. Applying sequences of such functions, f_n^N, to the source subsequence a_n^N we attain a set of target subsequences:

$$\mathrm{F}\left(a_n^N\right)=\left\{\left(f_1^k\left(a_n^N\right), f_2^k\left(a_n^N\right), \dots, f_{L_k}^k\left(a_n^N\right)\right)\in B^{L_k}\cup\varnothing\ ,\quad 1\leqslant k\leqslant K_{\mathrm{F}}\right\} \tag{2}$$

Here L_k is length of *k*-th target subsequence and the number of the target subsequences is K_F. Introduced correspondences (2) form a set $\mathbf{F}$.

Now we define an operation $\otimes$ that concatenates over the sets produced by F and G taken from **F** as all possible combinations of target sequences generated by F followed by G:

$$\mathrm{F}\otimes\mathrm{G}=\left\{\left(f_1^u, f_2^u, \dots, f_{L_u}^u, g_1^v, g_2^v, \dots, g_{L_v}^v\right),\quad 1\leqslant u\leqslant K_{\mathrm{F}},\ 1\leqslant v\leqslant K_{\mathrm{G}}\right\}. \tag{3}$$

Additionally, we assume that the connection result is empty if at least one of F or G is empty.

Further, we specify ordered correspondences (2) and accomplish them with additional parameters attaining a set:

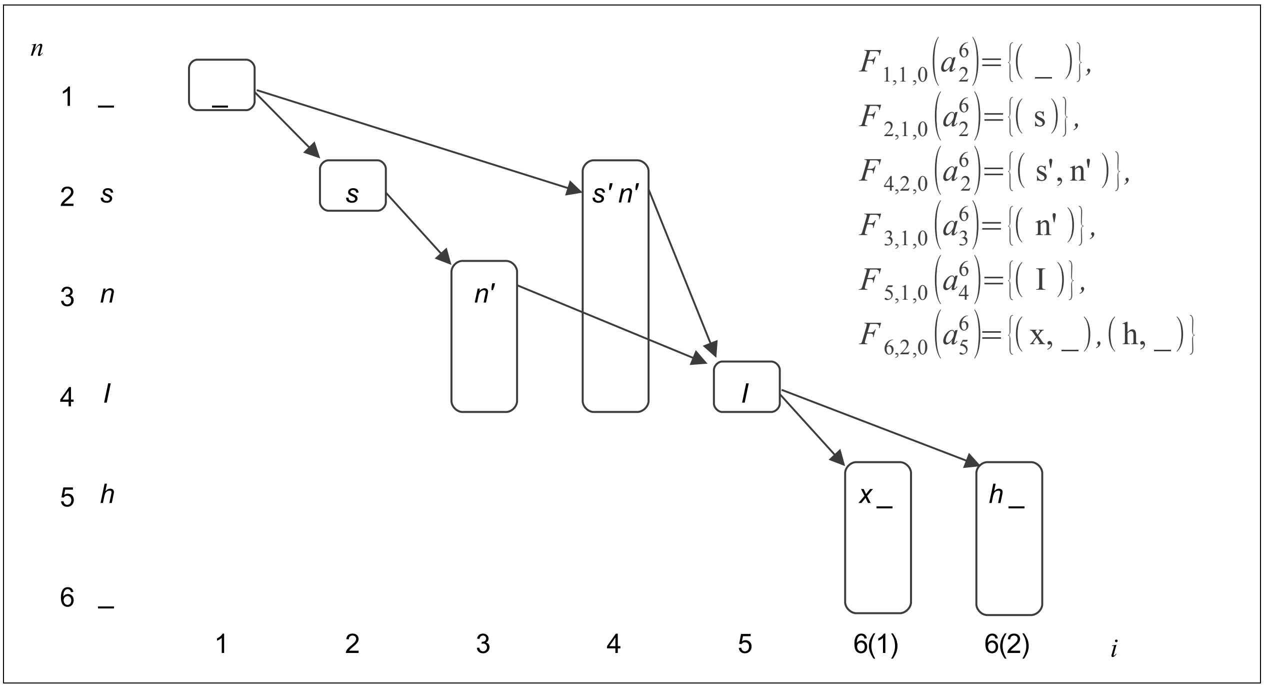
n
1 _
2 s
3 n
4 I
5 h
6 _
_
s
s'n'
n'
I
x_
h_
1
2
3
4
5
6(1)
6(2)
i
$F_{1,1,0}\left(a_2^6\right)=\{(\ \_\ )\},$
$F_{2,1,0}\left(a_2^6\right)=\{(\ \mathrm{s})\},$
$F_{4,2,0}\left(a_2^6\right)=\{(\ \mathrm{s}',\mathrm{n}'\ )\},$
$F_{3,1,0}\left(a_3^6\right)=\{(\ \mathrm{n}'\ )\},$
$F_{5,1,0}\left(a_4^6\right)=\{(\ \mathrm{I}\ )\},$
$F_{6,2,0}\left(a_5^6\right)=\{(\mathrm{x},\ \_),(\mathrm{h},\ \_)\}$

Figure 1: *Graph of multidecision grapheme-to-phoneme conversion for Ukrainian word "snih".*

$$\tilde{\mathbf{F}}=\left(\mathrm{F}_{i,d_i,\delta_i}\right),\ \mathrm{F}\in\mathbf{F}\ \ 1\leqslant i\leqslant|\tilde{\mathbf{F}}|,\ , \qquad (4)$$
$$0<d_i,\ \delta_i=\{0,\ 1\}$$

where d_i we call an analysis step and δ_i is an exclusivity condition for the i-th correspondence. Within these parameters we construct restricted connections

$$\underset{i,n}{\otimes}\mathrm{F}_{i,d_i,\delta_i}\left(a_n^N\right),\ 1\leqslant i\leqslant|\tilde{\mathbf{F}}|\ ,\ 1\leqslant n\leqslant N\ \ . \qquad (5)$$

Firstly we assume that (5) has already been evaluated for certain index sets J and M, which are ordered, and obtained

$$\mathrm{G}_{J,M}=\underset{u\in J,v\in M}{\otimes}\mathrm{F}_{u,d_u,\delta_u}\left(a_v^N\right)\ . \qquad (6)$$

Let j and m be the last elements J and M respectively. Then connecting the next correspondence, $\mathrm{F}_{i,d_i,\delta_i}\left(a_n^N\right)$, we proceed in accordance to (3), if the following conditions are met:

$$\begin{cases} m+d_i=n\ ; \\ \left[\begin{array}{l}\delta_r,\ 1\leqslant r\leqslant i\ ; \\ \underset{u\in J,v\in M}{\otimes}\mathrm{F}_{u,d_u,\delta_u}\left(a_n^N\right)\otimes\mathrm{F}_{r,d_r,\delta_r}\neq\varnothing\ ,\ 1\leqslant r\leqslant i\ ,\ \text{if}\ \ \delta_i=1\ .\end{array}\right.\end{cases}$$
(7)

Otherwise the connection is not applicable.

By means of expression (5) we can generate target sequences proceeding from a source sequence of symbols.

We illustrate this process on the graph in Figure 1. The Ukrainian word **сніг** – "snih" (snow) is accomplished with a word boundary symbol "_". We also uppercased a stressed vowel **i**.

Thus we have a sequence of six symbols a_1^N = ("_", "*s*", "*n*", "*I*", "*h*", "_"), $N=6$. All applicable correspondences $\mathrm{F}_{i,d_i,\delta_i}\left(a_n^N\right),\ 1\leqslant n\leqslant N$ are shown in the graph. Moving alongside the arrows we generate expressions of the form (5) receiving the following phoneme sequences or phoneme texts:

$$\_\ \mathrm{s}\,n'I\,x\ \_;\ \_\ \mathrm{s}\,n'I\,h\ \_;\ \_\ \mathrm{s}'\,n'I\,x\ \_;\ \_\ \mathrm{s}'\,n'I\,h\ \_.$$

Here the phoneme "*h*" is voiced and "*x*" is its voiceless pair.

On practice we do not need to consider the entire subsequence a_n^N . Normally, we narrow the context to $a_n^{n-1+T_F}$, where $T_\mathrm{F}\geqslant 1$ depends on the specific correspondence (2). In Figure 1 hight of rectangles corresponds to the context widths.

The expert can specify parameters of correspondences (4) in tabular way that is shown in the system description section. Note that we may apply the same or another set of correspondences to target sequences multiple times, therefore we introduce multiple levels for the conversion procedure. This allows for simplifying the parameter specification, which is important for cases when pronunciations are far from spelling. The other benefit of introduced levels is the possibility to convert numbers, symbols and abbreviations to their orthographic presentation within the same algorithm.

Moving from pronunciation to spellings we should track hypothetical letter sequences to detect word boundaries and remove non-applicable letter sequence hypothesis by referring a lexicon.

2.Word stress pointing procedure

For Ukrainian, normally, word stress position is a necessary hint for grapheme-to-phoneme conversion applied in text-to-speech systems. Stress position is irregular, it can change even within forms of the same word. Anyway, it is not acceptable to point stresses manually for the entire lexicon. Therefore, we propose a word stress prediction procedure based on the known lexicon and a text corpus.

We consider all possible segmentations S for the word with unknown stress. The i-th segment of S

$$S_i=\left(q(S_i,1),\ldots,q(S_i,j),\ldots,q(i,L(S_i))\right) \qquad (8)$$

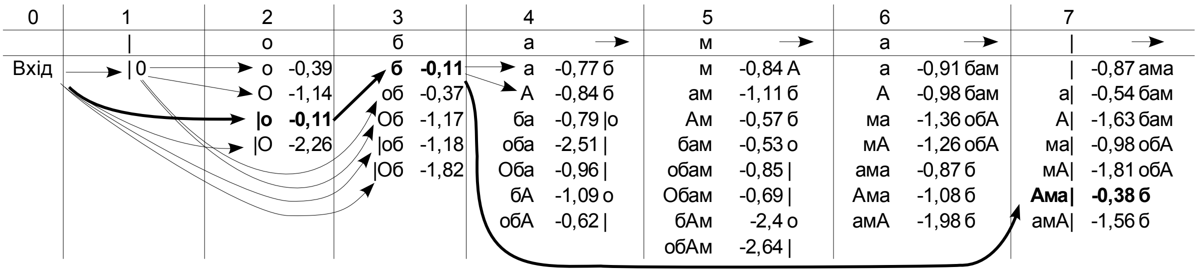

Figure 2. Stress prediction for an out-of-vocabulary word “обама” (Obama).

has length of $L(S_i)$. Here $q(S_i, j)$ is the *j*-th item (a character or a phoneme) within the *i*-th segment of *S*. Now we introduce a vector θ_L indicating the stress value (e.g. 0, 1, 2) for L items. We can estimate the probability of stress position given a segment S_i:

$$P\left(\theta_{L(S_i)} \mid S_i\right) \approx \frac{c\left(S_i, \theta_{L(S_i)}\right)}{c\left(S_i\right)} \tag{9}$$

where $c\left(S_i, \theta_{L(S_i)}\right)$ is count of segments S_i with stress position defined by a stress indication vector $\theta_{L(S_i)}$ and $c\left(S_i\right)$ is the number of total occurrence of a segment S_i. All counts are taken from the text corpus but the words not included in stress vocabulary.

Finally we search through all valid segmentations S and stress positions θ_S by the following the expression:

$$\underset{S,\theta_S}{\operatorname{argmax}} \prod_{S_i, \theta_{L(S_i)}} P\left(\theta_{L(S_i)} \mid S_i\right). \tag{10}$$

We can construct a dynamic programming graph where finding the shortest trajectory is equivalent to the search (10). Memorizing N prospective arrows to nodes of the graph we can extract N-best word stress positions supplemented with the criteria showing a confidence level for each solution.

In Figure 2 an example of 1-best stress prediction is shown for a proper name “Obama” missing from the basic Ukrainian vocabulary. The word is represented as concatenation of all valid character segments where the largest segment length is not longer than 4. Each input character introduces a set of valid segments. Potentially optimal arcs are either shown or coded with the name of a previous node. Partial criteria are log probability based. The optimal path, respective nodes and criteria are bold.

Stress error rate estimation in not as obvious procedure, since in specific cases it is unclear what is a mistake. E.g. the stress is predicted in erroneous words but if the prediction is mistaken why should we be as strict? Anyway, preliminary experiments exposed error level between 5 and 10% relatively to the vocabulary.

3.The system for multilevel multidecision text-to-pronunciation conversion

The developed text-to-pronunciation system consist of three modules: (a) grapheme extractor, (b) word stress pointer and (c) grapheme-to-phoneme converter (Figure 2).

Initially, the grapheme extractor (a) detects boundaries of words and complex word components, converts numbers, symbols and abbreviations to words.

Module (b) looks for word stress in the dictionary and, if it fails, tries to predict word stress positions in accordance to (8)-(10).

Thus we attain the input for the basic grapheme-to-phoneme converter (c) that is orthographic text containing solely alphabetic characters of the language accomplished with word/morpheme boundary and stress mark. The expert specify grapheme-to-phoneme correspondences (4) indicating source and target subsequences, step and exclusivity parameter. For convenience, we provide the generalized compact form for sequence specification illustrated in Table 1.

Only about 30 rules cover the literary pronunciation for Ukrainian. GTP conversion for spontaneous speech is a complicated case. Therefore, we introduce additional model levels allowing the expert to introduce pronunciations less resembled to spelling, still working with relatively simple correspondences.

The common pronunciation dictionary was created for all speakers by the basic set of regular rules. Furthermore, speakers were prescribed to specific groups correspondingly to their pronunciation peculiarities. Each group has the respective set of rules contributing irregularities to the basic set of rules.

Table 1. *Specification of grapheme-to-phoneme correspondences.*

Source subsequence	Target sub-sequence	Step size	Explanations
[зсц] [жшч]	[жшч]	1	**з**, **с**, **ц** before **ж**, **ш**, **ч** correspond with **ж**, **ш**, **ч**
[тс] [дтзснц] [іІєюяЄЮЯ ь]	т'	1	**т** and **с** before palatalisable **д**, **т**, **з**, **с**, **н**, **ц** go palatalized
с т [лн]	с	2	**т** between **с** and **л** or **н** is eliminated

Analysis of pronunciation for a big amount of speakers shows that no one follows thoroughly the regular pronunciation rules. Firstly it concerns to regularly forbidden regressive invoicing assimilation in pair of phones “voiced+unvoiced” and consonant devocalization before a pause: **тобто → т О п т о (tobto → t O p t o)**; **підтримати → п' і т т р И м а т и (p'idtrymaty → p' i t t r Y m a t y)**; **робив → р о б И ф (robyv → r o b Y f)**. Speakers with such peculiarities were selected to the separate group.

Many other distinctive features of pronunciation of different speakers have been detected as well. These are such features: reduction of the terminations of some words (adjectives, verbs) in continuous speech, partial vowel reduction, non-palatalized pronunciation of palatalized consonants: **шановний → ш а н О в н и (shanovnyj → sh a n O v n y)**; **доброго → д О б р о (dobroho → d O b r o)**; **робити → р а б И т и (robyty → r a b Y t y)**; **синього → с И н о г о (syn'oho → s Y n o h o)**.

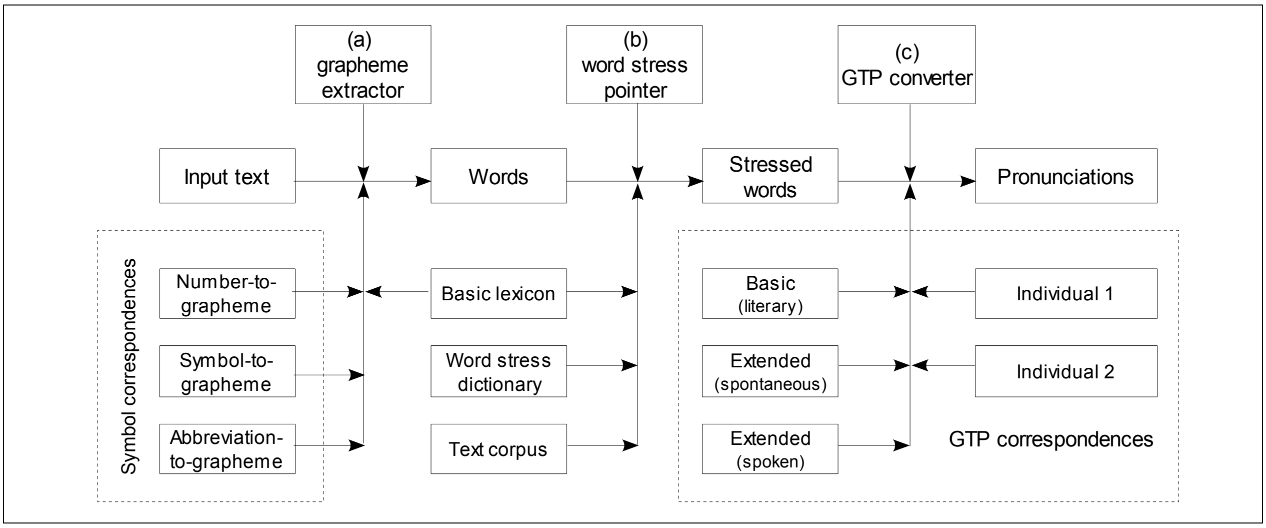

Figure 3. *Text-to-pronunciation conversion system structure*

For some words (specific parts of speech, words with different accents) a few variants of transcriptions are introduced. Word accent is set on different syllables (if different variants of reading of such words are permissible in the language) or without any accentuation at all: **коли → к о л И; к О л и; к о л и (koly → k o l Y; k O l y; k o l y)**.

Such tendencies are implemented by changing of transition rules from one sequence of characters to the other and by expansion of existent rules.

All rules of the individualized modification of transcriptions can be divided into few groups [6].

Changes of sounds which depend on general phonetic conditions like position in syllable or word, accentuation property and others:

- reduction of unstressed *е (e)*, *и (y)* and *о (o)* to $е^{и}$ *(*e^{y}*)*, $и^{е}$ *(*y^{e}*)*, $о^{у}$ *(*o^{u}*)*, weak pronunciation *о* as *а* in unstressed position in the word, reduction of unstressed vowels till complete disappearance;
- pronunciation of final voiced consonants as voiceless ones;
- reduction in terminal parts of words during the pronouncing process (disappearance of consonant in word completions *-ого*, *-их*, *-ий*, *-іх*, *-ій*, -ії, *-ої*, *-еї*, *-ою*, *-єю*, *-ити (-oho, -ykh, -yj, -ikh, - ij, -iji, -oji, -eji, -oju, - jeju, -yty)* and others);
- disappearance of an unstressed vowel in word completions *–ою*, *-ею*, *-єю (-oju, -eju, - jeju)* etc.

Qualitative and quantitative changes of adjacent phones:

- complete regressive voiceless assimilation in combinations „voiced+unvoiced" on boundaries of any morphemes inside words and on boundaries of words;
- palatalization of sibilant, labial and velar consonants in certain contexts;
- pronouncing of long and doubled consonants as one phoneme, pronouncing of two running vowels as one phoneme;
- incomplete simplification in groups of consonants etc.

It take 1 to 2 weeks to train an expert to specify the correspondences. As reported in [6] speech recognition relative WER decreased about 5% for individualized pronunciation dictionaries.

4.Conclusions

The proposed model allows expert to describe regularities for conversions between text and its pronunciations in a compact and convenient way. Numbers, symbols and abbreviations can be converted to their textual presentation by feeding the same algorithm with another sets of symbol sequence correspondences.

Introduced levels allow for constructing the data for languages for which pronunciations go far from spelling. Thus, for Russian language the literary pronunciation needs 7 levels and about 200 rules and we expect results for English.

Currently we do not estimate scores for the extracted target symbol sequences. Estimation of respective probabilities is possible by referring to recognition result analysis.

Proposed stress pointing prediction procedure may be driven from either grapheme or phoneme data. It is not limited to neither language nor level of speech patterns hierarchy. The extracted optimal segmentation defines an alphabet of sub-word units that express both prosodic and morphologic nature of words. Segment context introduction is a subject of further research.

5.References

[1] T. Vintsiuk, T. Lyudovyk, M. Sazhok, R. Selyukh. Automatical Ukrainian text-to-speech conversion based on phoneme-triphone model with natural speech signal using. Proc. of the 6th All-Ukrainian Int. Conf. on Signal/Image Processing and Pattern Recognition "UkrObraz'2002", Kyiv, 2002, pp 79-84, in Ukrainian.

[2] Vintsiuk T.K., Analysis, recognition and understanding of speech signals. Kyiv: Naukova Dumka, 1987, 264 p., in Russian

[3] M. Gales and S. Young. "The Application of Hidden Markov Models in Speech Recognition." Foundations and Trends in Signal Processing, 2007, 1(3): 195-304.

[4] Taras Vintsiuk, Mykola Sazhok. Multi-Level Multi-Decision Models for ASR // Proceedings of the 10th Int. Conference on Speech and Computer – SpeCom'2005, Patras, 2005, pp. 69-76.

[5] Bisani, M., Ney, H. Joint-sequence models for grapheme-to-phoneme conversion // Journal Speech Communication, 50: 434-451, Elsevier, 2008.

[6] V. Pylypenko, V. Robeiko. Experimental System of Computerized Stenographer for Ukrainian Speech. Proc. of the 13th International Conference SPECOM'2009, St.-Petersburg, RF, 2009. Pp. 67—70.