Unit Selection Speech Synthesis Using Phonetic-Prosodic Description of Speech Databases

Tetyana Lyudovyk, Mykola Sazhok

International Research/Training Center for Information Technologies and Systems, Kyiv, Ukraine
{tetyana_lyudovyk, mykola}@uasoiro.org.ua

Abstract

This paper describes an approach to speech synthesis based on using speech databases at different stages of TTS process. Speech database units are phones in different segmental and prosodic contexts. Pitch synchronous segmentation and labeling of databases allows storing both segmental and prosodic information.

Phonetic-prosodic annotations of speech databases are involved in off-line training of the linguistic processor. The automatic transcriptor, duration and intonation modules are trained to model the speech characteristics of different persons and thus to generate different target specifications of one and the same input text during the synthesis stage. A target specification is a detailed phonetic-prosodic transcription used by the unit selection module.

The unit selection algorithm is based on criteria derived from categories of phonetic-prosodic annotations of speech databases and works without spectral matching. The output of the unit selection module is an acoustic phonetic-prosodic transcription which is used by the acoustic processor to generate a speech wave.

Two non-professional speaker databases with different speaking styles have been created and tested.

1. Introduction

Speech databases play a great role in concatenative synthesizers. Extending a database size and coverage we increase the probability of finding speech units with specified properties, e.g context, duration and F0 contour. Consequently we decrease the need to modify the speech signal. As a result we obtain the synthesized speech which is more natural and of higher quality.

Speech databases are individual and reflect the characteristic pronunciation properties of speakers. Using this information ensures that a speaker will be recognized by the synthesized speech. The novelty of the suggested approach consists in using the database phonetic-prosodic annotation not only for on-line unit selection but also for off-line training of all modules of the linguistic processor. The trained linguistic processor is able to generate speaker-specific target specifications of input texts. The main idea is that the closer the target specification is to the database speaker pronunciation, the more precise and correct unit selection and overall synthesis quality will be achieved.

The overall approach is time domain oriented. It concerns database collection, segmentation and labeling, as well as unit selection criteria used and signal processing techniques applied (optional).

The described approach is realized in the experimental Ukrainian TTS system. Two speech databases with male voices (8000 and 12000 units) are developed and a speech database with a female voice is under developement.

2. Components of the Ukrainian TTS system

The experimental Ukrainian TTS system is composed of the traditional modules [1], [2], [3]: speech database, linguistic processor, acoustic processor, and unit selection module [3], [4], [5], which is indispensable in synthesizers using large speech databases. The overall system architecture is shown in Figure 1.

Speech database plays the main role. The information it contains (detailed annotation of acoustic units corresponding to phones in context) is used by all other modules of the synthesizer.

The pre-tuning is applied offline to all modules of the linguistic processor (text normalization, accentuation and phrasing, phonetic transcription, assignment of duration and F0 contour.

The annotation of the speech database is used also as the main information source for unit selection during the synthesis stage. The criteria used by the unit selection algorithm are based on the annotation categories: left and right context of a unit, its duration and F0, as well as the sequence of units in the speech corpus preserved in the database.

Speech signal segments stored in the database are used by the acoustic processor for generating the resulting speech waveform corresponding to the input text [6].

3. Speech database

The quality of the synthesized speech depends on the size and the coverage of the speech database. Another important factor is the level of detail of the description which is used to represent the speech units.

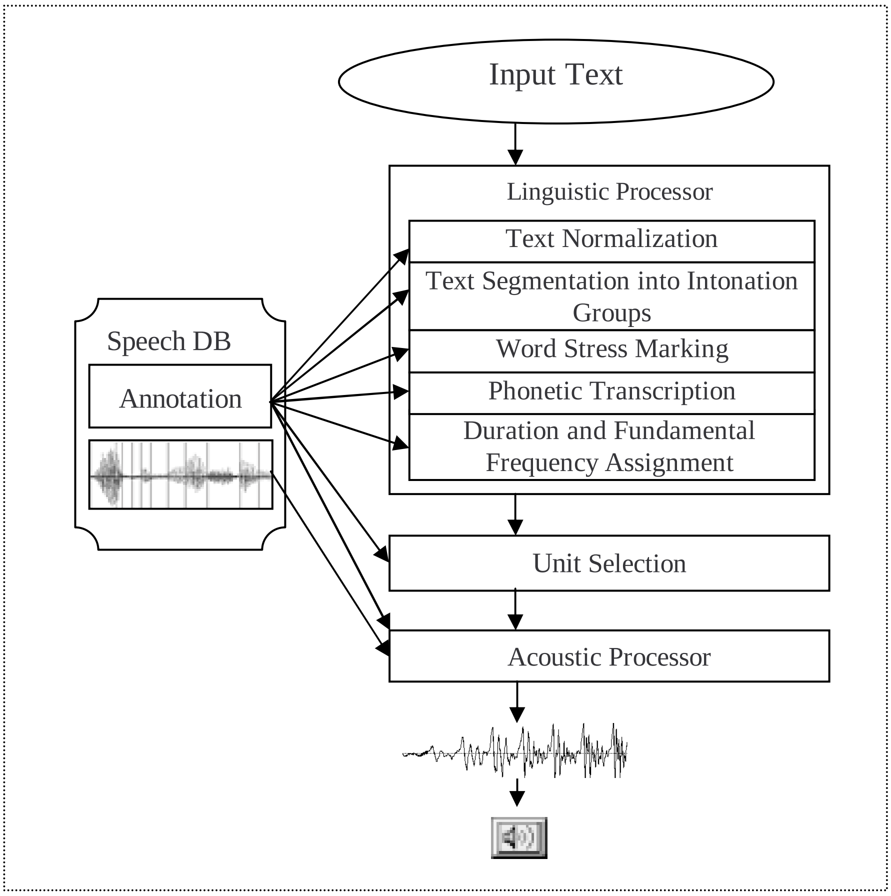

Figure 1. Block diagram of the Ukrainian TTS system

3.1. Database design

Speech database units are phones (phonemes and distinct allophones) in context. Phones have been chosen because they are the main segmental phonetic units and appear as main elements of target specifications generated for input texts.

We distinguish 58 phones in Ukrainian: 6 stressed and 6 unstressed vowels, 22 non-palatalized consonants, 23 palatalized consonants, and pause. So, stress is present explicitly in the phoneset, which is typically the case for Eastern Slavic language synthesizers [1], [2].

Different speech corpora are recorded using non-professional speakers reading aloud texts and isolated sentences. Recordings are made at 22 kHz sampling rate.

Multiple instances of each phone differing in duration and pitch are stored.

Phones are collected in the order they occur in training sentences. This information is used by unit selection algorithm instead of continuity distance measurement.

The size of speech corpora used for database creation is moderate, which allows to manually correct its labeling and segmentation. As a result the annotations of speech databases are reliable and reflect the slightest peculiarities of speaker pronunciation. This fact allows for using these speech databases not only for text-to-speech synthesis, but also for creation of linguistic models (e.g. prosody model) and for different kinds of experimental-phonetic research.

3.2. Database collection and segmentation

The collection of a database is done under expert control with possibility to see and hear every unit instance thanks to interactive visualization tools [7].

The segmentation of speech corpus into phones is done manually, because it requires correcting of automatically obtained transcription of training texts [8] to how a particular speaker reads these training texts (different accentuation, presence/absence of assimilation, reduction, insertion of extra pauses etc.).

The second stage of segmentation, into pitch periods, is carried out automatically [6], [9], but it also requires an expert supervision.

Pitch marks are positioned on zero crossing points before the first negative or positive peak of the pitch period, depending on speech material. Each pitch period is visually checked and possibly corrected. This allows to obtain the accurate description of F0 movement, reflecting the details of microprosody.

The database is pruned by removing the majority of pauses, as well as “poorly” pronounced phones.

3.3. Database unit description

Each phone in context may occur several times in the database, so untis need identifiers. Segmentation provides information about unit duration and F0 trajectory. Only voiced units are segmented into pitch periods.

A fragment of one annotation is shown in Figure 2. Each instance of a unit in the database is labeled with:

- unit instance identifier;
- three-part unit name (preceding, current, succeeding phone names);
- unit instance duration in ms;
- number of pitch periods (only for voiced units);
- average pitch period length (only for voiced units);
- first pitch period length (only for voiced units);
- medium pitch period length (only for voiced units);
- last pitch period length (only for voiced units).

2725	a - l - Y	64.08	10	6.39	5.90	6.62	6.17
2726	l - Y - s	80.32	15	5.35	6.08	5.31	5.03
2727	Y - s - a	115.51					
2728	s - a - h	68.21	11	6.17	5.53	6.17	6.94
2729	a - h - o	83.95	11	7.62	7.39	7.53	7.71
2730	h - o - l	55.42	7	7.89	7.53	7.98	8.30
2731	o - l - o	40.18	5	8.03	7.89	8.07	8.12
2732	l - o - v	62.54	8	7.80	6.98	8.89	8.16
2733	o - v - A	61.63	8	7.66	8.07	7.62	7.30
2734	v - A - #	159.41	27	5.90	7.17	5.90	4.90

Figure 2. A fragment of the speech database annotation.

As a result, a database unit description comprises both segmental and prosodic characteristics of speech.

4. Linguistic processor training

The goal of the linguistic processor in a unit selection concatenative speech synthesizer is to generate the target specification of an input orthographic text. Typically, linguistic processors are independent of the database [3], [5], [10], [11]. Our goal, on the contrary, is to produce such a target specification that would model a speaker specific pronunciation (e.g. different word accentuation, different pitch range, different pitch contours for continuation rise, enumeration etc.). We believe that making target specification close to database pronunciation would help significantly with unit selection as well as with naturalness and quality of the synthesized speech.

Linguistic processor components use two types of data: speaker independent and speaker dependent. Speaker independent data are the phoneset and phones features. Speaker dependent data are:

- dictionaries for text normalization and word stress marking;
- maximum length of an intonation group (in phonetic words);
- rules for phonetic assimilation and reduction;
- average durations of phones and duration lengthening and shortening coefficients;
- intonation contour inventories.

Let us examine some of the speaker dependent data taking as an example two speech databases with male voices. The maximum length of an intonation group is correspondingly 7 and 8 phonetic words. The average durations of stressed vowels are:

- [A]: 117 ms and 95 ms;
- [O]: 102 ms and 84 ms;
- [E]: 104 ms and 84 ms;
- [I]: 88 ms and 68 ms;
- [Y]: 87 ms and 69 ms;
- [U]: 94 ms and 78 ms.

The same speakers use different F0 contours for non-finality. The fundamental frequency assignment module produces speaker dependent target contours, as shown in Figure 3.

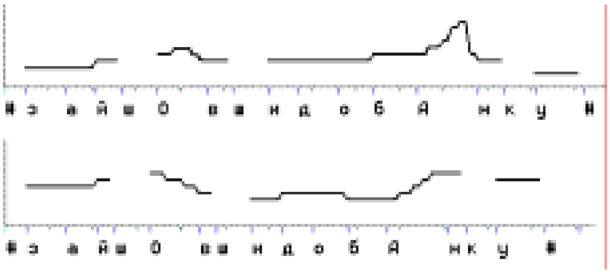

Figure 3. Speaker dependent target F0 curves in case of the same intonation group with non-finality (in transliteration): “Zajshovshy do banku“ (“entered the bank”).

At present the fully automated training is developed for phone duration prediction. The phone duration prediction is based on the phone average duration and a set of duration coefficients. Each phone of each speaker has its own average duration and its own set of duration coefficients. Moreover, these data depend on the speaking style. For synthesizing speech in different speaking styles we suggest to use different speech databases recorded from one donor-speaker.

5. Target specification

The target specification provided by the linguistic processor is the phonetic-prosodic transcription of the input text. To obtain the target specification the following procedures are involved: text normalization, and segmentation into intonation groups and intonation type assignment, word stress marking, grapheme-to-phone conversion, segmenting of intonation groups into accent groups, computing F0 curves and duration of phones, conversion of the sequence of F0 values into the sequence of pitch period lengths. The resulting target phonetic-prosodic transcription is used during unit selection.

The target phonetic-prosodic transcription for the same intonation group “Zajshovshy do banku“ is present in Figure 4. Phones-in-context are followed by sequences of pitch period lengths for voiced phones and simply by duration for unvoiced phones (in ms).

#-#-z	50
#-z-a	8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0
z-a-j	8.0 8.0 7.9 7.9 7.9 7.8 7.8 7.8 7.7
a-j-sh	7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7
j-sh-O	80
sh-O-v	7.4 7.4 7.3 7.3 7.2 7.2 7.2 7.2 7.3 7.3 7.4 7.5 7.5 7.6 7.7
O-v-sh	7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7
v-sh-y	80
sh-y-d	7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7
y-d-o	7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7
d-o-b	7.7 7.7 7.6 7.6 7.6 7.6 7.5 7.5 7.5 7.4 7.4
o-b-A	7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4
b-A-n	7.4 7.3 7.3 7.2 7.1 7.1 7.0 7.0 6.9 6.8 6.6 6.5 6.4 6.4 6.3 6.6 6.9 7.3 7.4 7.5 7.7
A-n-k	7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7
n-k-u	70
k-u-#	8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3 8.3
u-#-#	80

Figure 4. Target specification for the intonation group “Zajshovshy do banku”.

6. Unit selection

The unit selection algorithm is based on phonetic and prosodic criteria. It uses:

- target phonetic-prosodic transcription of the input text provided by the linguistic processor;
- phonetic-prosodic annotation of the speech database;
- tables of phonetic-acoustic distances between Ukrainian phones;
- phonetically motivated criteria of selection.

The main criterion of unit selection algorithm according to phoneme-triphone model [12] is contextual identity of target and candidate units. Left and right neighbors of each target unit are taken into account. The algorithm first searches the database for candidate units that match the target unit along with its left and right contexts. If there are such units in the database, the selection continues in two different ways for periodic and non-periodic units.

For periodic units the following prosodic selection criteria are used:

- difference between target and candidate units in average pitch period lengths;
- difference between target and candidate units in pitch periods number.

For non-periodic phones the difference between target and candidate units durations is used as the selection criterion.

One more criterion is the immediate vicinity of units in the database. The selection of consecutive phones is encouraged.

If the database does not contain units with left and right segmental context equal to the target segmental context, tables that specify the phonetic-acoustic distances between Ukrainian phones are used to search for unit with context similar to the target unit context.

The result of unit selection procedure is a specification for the acoustic processor. This specification is called acoustic phonetic-prosodic transcription and it differs from the target phonetic-prosodic transcription by:

- the indication of database units identifiers;
- the indication for periodic units what pitch periods in what quantity (for unit duration changing) and with what length (for unit F0 changing) should be taken for the concatenation.

The resulting unit selection specification for the intonation group presented in sections 4 and 5 is shown in Figure 5. Note that for several units some lengthening or shortening is required (certain pitch periods must be repeated or deleted). Units neighboring in the database are presented in bold. Database units with context not identical to the target unit context are presented in italic.

1271	#-#-z	0 50
2162	#-z-a	(1 8.0) (2 8.0) (4 8.0) (5 8.0) (6 8.0) (7 8.0) (9 8.0)(10 8.0)(11 8.0)(12 8.0)(14 8.0)(15 8.0)
4240	z-a-j	(1 8.0)(2 8.0)(4 7.9)(5 7.9)(7 7.9)(8 7.8) (10 7.8)(11 7.8)(13 7.7)
4164	a-j-sh	(1 7.7)(2 7.7)(3 7.7)(3 7.7)(4 7.7)(5 7.7) (6 7.7)(6 7.7)
4165	j-sh-O	0 80
4166	sh-O-v	(1 7.4)(2 7.4)(4 7.3)(5 7.3)(7 7.2)(8 7.2) (10 7.2)(11 7.2)(13 7.3)(14 7.3)(16 7.4)(17 7.5) (19 7.5)(20 7.6)(22 7.7)
1004	***O-v-t***	(1 7.7)(2 7.7)(2 7.7)(3 7.7)(4 7.7)(4 7.7) (5 7.7)(5 7.7)
4759	v-sh-y	0 80
2065	***k-y-d***	(1 7.7)(2 7.7)(2 7.7)(3 7.7)(4 7.7) (4 7.7)(5 7.7)(6 7.7)(7 7.7)(7 7.7)

2590	y-d-o	(1 7.7)(2 7.7)(3 7.7)(3 7.7)(4 7.7)(5 7.7) (6 7.7)(7 7.7)(8 7.7)(8 7.7)(9 7.7)(10 7.7)
974	***d-o-d***	(1 7.7)(2 7.7)(3 7.6)(4 7.6)(5 7.6) (6 7.6)(7 7.5)(8 7.5)(9 7.5)(10 7.4)(11 7.4)
2156	o-b-A	(1 7.4)(2 7.4)(2 7.4)(3 7.4)(4 7.4) (5 7.4)(5 7.4)(6 7.4)(7 7.4)(7 7.4)(8 7.4)(9 7.4) (10 7.4)(10 7.4)
5594	b-A-n	(1 7.4)(2 7.3)(3 7.3)(4 7.2)(5 7.1)(6 7.1) (7 7.0)(8 7.0)(9 6.9)(10 6.8)(11 6.6)(11 6.5) (12 6.4)(13 6.4)(14 6.3)(15 6.6)(16 6.9)(17 7.3) (18 7.4)(19 7.5)(20 7.7)
5595	A-n-k	(1 7.7)(2 7.7)(4 7.7)(5 7.7)(7 7.7)(8 7.7) (9 7.7)(11 7.7)
3140	n-k-u	0 70
1168	k-u-#	(1 8.3)(2 8.3)(3 8.3)(4 8.3)(5 8.3) (6 8.3)(7 8.3)(7 8.3)(8 8.3)(9 8.3)(10 8.3) (11 8.3)(12 8.3)(13 8.3)
539	u-#-#	0 80

Figure 5*:* Unit selection module output for the intonation group “Zajshovshy do banku”.

7. Unit concatenation

After the appropriate sequence of units is selected from the database, the next step is to concatenate the segments. This is done by the acoustic processor.

To obtain the prosody specified in the target, it is usually necessary to modify the duration and fundamental frequency.

Our time domain approach ensures prosodic modification of selected units via signal processing. The prosodic modification of a database unit consists in repeating or omitting its pitch periods according to the target acoustic phonetic-prosodic transcription, as well as in lengthening or shortening of selected pitch periods applying the linear prediction model [12].

The speech signal processing inevitably distortes the real speech segments and thus degrades the naturalness of synthesized speech. The size of the speech database is important, because among a great number of segmentally and contextually identical phones it is more probably to find a phone with appropriate prosodic characteristics.

At present two versions of speech generation are available. The first version modifies each selected pitch period of selected unit according to target specification. The second version does not modify selected units at all; that is, they are taken from the database with their real duration and pitch.

Consistent pitch marking of speech data and detailed target specification of prosody allow to avoid phase and pitch mismatches at speech generation stage [13], while taking into account coarticulation reflected in contexts allows to minimize spectral mismatch.

8. Results

Three allophone-based databases with three different donor speakers have been created:

- artificial words with careful monotone rather slow pronunciation (712 items);
- real speech, sentences from stories and novels (12057 items);
- real speech, radio-news reports (8433 items).

A listening test was conducted to evaluate, how the speech recordings quality and the signal processing which modifies the prosody influence the intelligibility and the naturalness of synthesized speech.

Two texts differing in style (a news report and a fiction story) were synthesized using the first and the second databases under following combinations:

A) with small database, without unit selection, with prosodic modification;

B) with medium size database (5873 items), with unit selection, with prosodic modification;

C) with medium size database (5873 items), with unit selection, without prosodic modification.

20 listeners, all native speakers of Ukrainian from different regions of Ukraine were asked to rate A, B, and C versions of synthesized speech according to intelligibility and naturalness.

The results are shown in Figure 6.

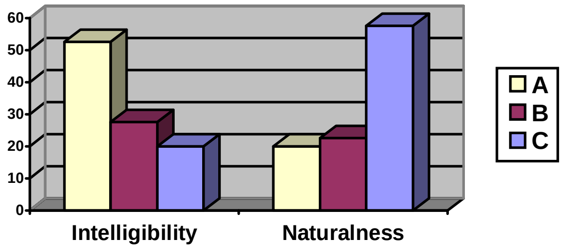

Figure 6*:* Listeners’ preference percentage for three tested synthesis technique combinations A, B, C.

The results of listening test indicate that the best intelligibility is achieved with small database. This is due to correct delivery of speech data during recordings. As for naturalness, the version C is preferred. This indicates that listeners tolerate spectral or prosodic mismatches if the quality of the synthetic speech closer approaches that of natural speech. Version B keeps an intermediate position.

A preliminary experiment related to fully automated segmentation and labeling of speech data was carried out with female speech recordings.

9. Future work

The future work will be concentrated on:

- selecting optimal speech data for databases;
- automatic creation of speech databases;
- pruning the size of databases by removing unit instances which are redundant or are never selected

due to their values out of established pitch or duration range;
- providing better representations for target, adding amplitude (energy) specification;
- automatic training of all modules of linguistic processor;
- refining unit selection criteria;
- elaborating new signal processing techniques;
- trying non-uniform units (demi-phones, phone clusters etc.).

10. Conclusions

The main goal of the work described was to develop and test a new speech technology based on:

- careful database design, segmentation and annotation;
- producing target specification close to database content in segmental and prosodic sense;
- simple unit selection algorithm.

Concatenative synthesis approach is able to retain the characteristics of the donor speaker in the synthesized speech. Speaker characteristics are reflected in the target specification.

Detailed segmental-prosodic annotation of database units and detailed phonetic-prosodic target specification allow selection criteria to be simple and easily manipulated.

The listening test showed that an important factor influencing the intelligibility of resulting synthesized speech is the database quality [14]. As for the naturalness, the strategy “select the best, modify the least” is confirmed to be true.

11. References

[1] Krivnova, O.F. “Automatic synthesis of Russian speech”, Proceedings of the XIV International Congress of Phonetic Sciences, Vol.1, 507–510, San Francisco, 1999.

[2] Лобанов Б.М., Карневская Е.Б., Левковская Т.В. “Синтезатор речи по тексту как компьютерное средство «клонирования» персонального голоса”, Тр. Международной конференции Диалог-2001, 265-272, М., 2001.

[3] Beutnagel, M., Conkie A., Schroeter, J., Stylianou, Y. and Syrdal, A. “The AT&T Next-Gen TTS System”, The Proceedings of the Joint Meeting of ASA, EAA, and DAGA, 18-24, Berlin, Germany, 1999.

[4] Hunt, A., Black, A. W. “Unit selection in a concatenative speech synthesis system using a large speech database”, Proceedings of the IEEE International Conference on Acoustics and Speech Processing, Vol. 1, 373-376, Munchen, Germany, 1996.

[5] Coorman, G., Fackrell, J., Rutten, P., Van Coile, B. “Segment Selection in the L&H RealSpeak Laboratory TTS System”, Proceedings of the International Conference on Spoken Language Processing, Vol. 2, 395-398, Beijing, China, 2000.

[6] Людовик Т.В., Сажок Н.Н. “Использование речевых баз данных большого объема при синтезе речи в системах искусственного интеллекта”, Проблемы управления и информатики, 6, 82-87, 2003.

[7] Sazhok, M. “Speech modelling virtual laboratory”, Proceedings of the International Summer School on Speech Processing, Recognition and Artificial Networks, 229–233, Vietri-sul-Mare, Italy, 1998.

[8] Vintsiuk, T. K., Liudovyk, T.V., Sazhok, M.M., “Phonetic Knowledge Base for Ukrainian”, Proceedings of the Second International Conference SPECOM’98, 179-192, St.Petersburg, Russia, 1998.

[9] Vintsiuk, T. K., “Optimal joint procedure for current pitch period discrimination and speech signal partition into quasi-periodic and non-periodic segments”, Proceedings of the First International Workshop on Text, Speech, Dialogue—TSD’98, 135-140, Brno, 1998.

[10] Conkie, A. “Robust unit selection system for speech synthesis”, The Proceedings of the Joint Meeting of ASA, EAA, and DAGA, 18-24, Berlin, Germany, 1999.

[11] Beutnagel, M., Conkie, A., and Syrdal, A. “Diphone synthesis using unit selection”, The Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, 185-190, Jenolan Caves, Australia, 1998.

[12] Тарас Вінцюк, Тетяна Людовик, Микола Сажок, Руслан Селюх. “Автоматичний озвучувач українських текстів на основі фонемно-трифонної моделі з використанням природного мовного сигналу”, Праці 6-ї Всеукраїнської міжнародної конференції “Оброблення сигналів і зображень та розпізнавання образів” – УкрОбраз’2002, Київ, 2002, с. 79–84.

[13] Bozkurt, B., Bagein, M., Dutoit, T. “From MBROLA to NU-MBROLA”, Proceedings of the 4th ISCA Speech Synthesis Workshop, 127-129, Pitlochry, Scotland, 2001.

[14] Möbius B. “Corpus-based speech synthesis: methods and challenges”, Arbeitspapiere des Institut fur Maschinelle Sprachverarbeitung, AIMS 6(4), 87-116, University of Stuttgart, 2000.