Speech Databases for Concatenative Speech Synthesis

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 a new technique for speech synthesis based on using speech databases at different stages of text-to-speech process. Speech databases are used for storing, selection and concatenation of speech segments. 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. 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.

The described approach is realized in the experimental Ukrainian TTS system. Several non-professional speaker databases with different speaking styles have been created and tested.

1. Introduction

Speech databases play a great role in concatenative synthesizers. The unit selection method of speech synthesis implies the development of large speech databases containing several thousand of speech units (allophones, diphones, half-phones). 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.

Pitch synchronous segmentation allows for obtaining a detailed description of speech data on both segmental and prosodic levels. It also reflects the differences in pronunciation by different speakers reading one and the same training text.

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).

2. Components of the Ukrainian TTS system

The experimental Ukrainian TTS system is composed of the following modules: speech database, linguistic processor, unit selection module, and acoustic processor. The overall system architecture is shown in Figure 1.

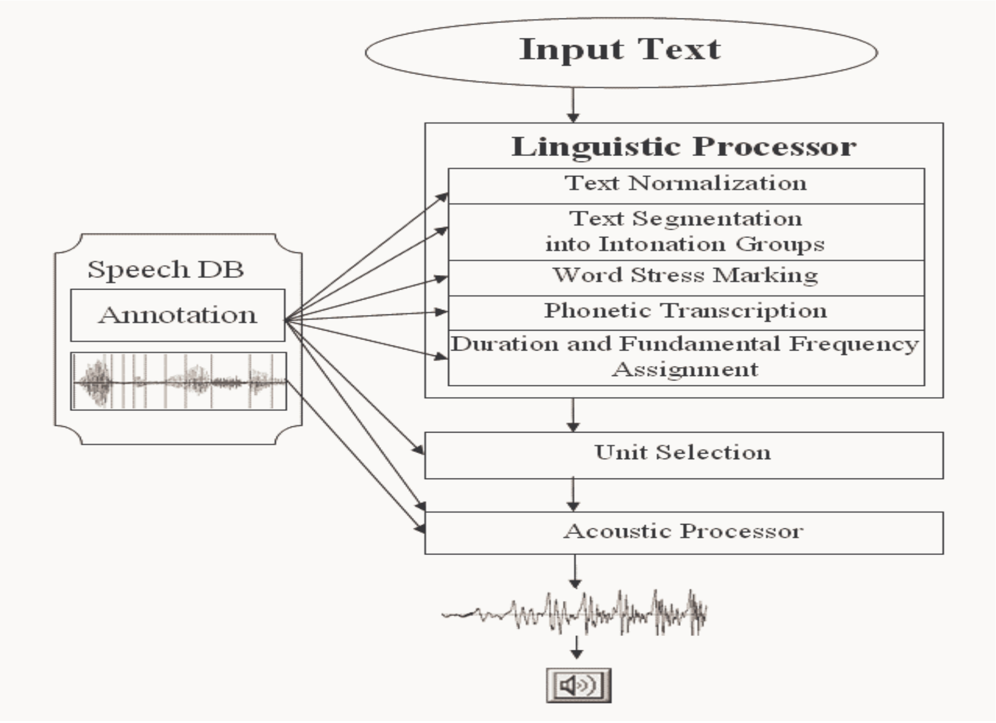

Figure 1. Block diagram of the Ukrainian TTS system

The main role is played by the speech database. It contains detailed annotation of acoustic units corresponding to phones in context, that is allophones. This information is used by all other modules of the synthesizer.

First, linguistic processor is trained to account for speaker specific phonetic and prosodic characteristics reflected in the database. This is done offline, now partly automatically, e.g. training of the module predicting phones duration.

Then during online speech synthesis the annotation of the database is used for unit selection, and along with the acoustic part of it, that is speech segments stored, is used by the acoustic processor for generating the resulting speech waveform [1].

3. Speech database collection and annotation

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 units description.

Several experimental speech databases, both male and female voices, have been collected and annotated. The speakers are non-professional speakers. One of the speakers read isolated sentences from stories and novels, another one read prompts used in interactive dialog, (e.g. “What language would you like to speak?”, “Please, say a command.”). Several voices have been captured from the Internet, these were news-reports. The recording sampling rate was 22 kHz.

The size of speech corpora used for database creation is moderate, not large, but this allows to manually correct the labeling and segmentation. As a result the annotations of speech databases are reliable and reflect the slightest peculiarities of speaker pronunciation.

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   5.90  6.71  6.71  6.62  6.62  6.44  6.53  6.26  6.08  6.17
2726  l-Y-s    80.32  15   6.08  5.71  5.80  5.62  5.53  5.44  5.53  5.31  5.03  5.03
                            4.99  4.99  5.03  5.12  5.03
2727  Y-s-a   115.51
2728  s-a-h    68.21  11   5.53  5.44  5.90  5.80  6.17  6.17  6.44  6.53  6.44  6.80
                           6.94
2729  a-h-o    83.95  11   7.39  7.17  7.35  7.62  7.80  7.53  7.89  7.80  7.80  7.85
                           7.71
2730  h-o-l    55.42   7   7.53  7.71  7.98  7.98  8.07  7.80  8.30
2731  o-l-o    40.18   5   7.89  8.07  8.07  7.98  8.12
2732  l-o-v    62.54   8   6.98  7.80  7.80  8.89  7.89  7.98  6.98  8.16
2733  o-v-A    61.63   8   8.07  7.98  7.98  7.62  7.62  7.76  7.26  7.30
2734  v-A-#   159.41  27   7.17  6.71  6.71  6.62  6.49  6.49  6.26  6.26  6.08  6.17
                            5.99  6.17  5.90  5.90  5.90  5.80  5.71  5.53  5.53  5.62
                            5.35  5.44  5.35  5.26  5.08  4.99  4.90
```

Figure 2. A fragment of the speech database annotation

The collection of a database is done under expert control. The second stage of segmentation, into pitch periods, is carried out automatically [2], but it also requires an expert supervision.

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

4. Linguistic processor training

Target specification for synthesis, that is segmental-prosodic transcription of an input orthographic text is provided by the linguistic processor.

Text analysis components (transcription, duration, and intonation prediction modules) 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; average and maximum length of an intonation group (in phonetic words); rules for phonetic assimilation and reduction; average duration of phones and duration lengthening and shortening coefficients; intonation contour inventories.

In Figure 3 different real stylized intonation contours of non-finality pronounced by the speaker S. are shown. These are non-finality intonation groups each composed of three accent groups. All these 16 intonation groups are composed of 3 accent groups, the last one is the nuclear one. The first point of each accent group represents the fundamental frequency at the middle of the first voiced phone of the accent group (voiced means vowels and voiced consonants). Then the next six points represent the fundamental frequency movement on the stressed vowel, that is the nucleus of this accent group, and the last point of the accent group represents the fundamental frequency at the middle of the last voiced phone of this accent group. The stylization is undertaken to avoid the representation of microprosody effects.

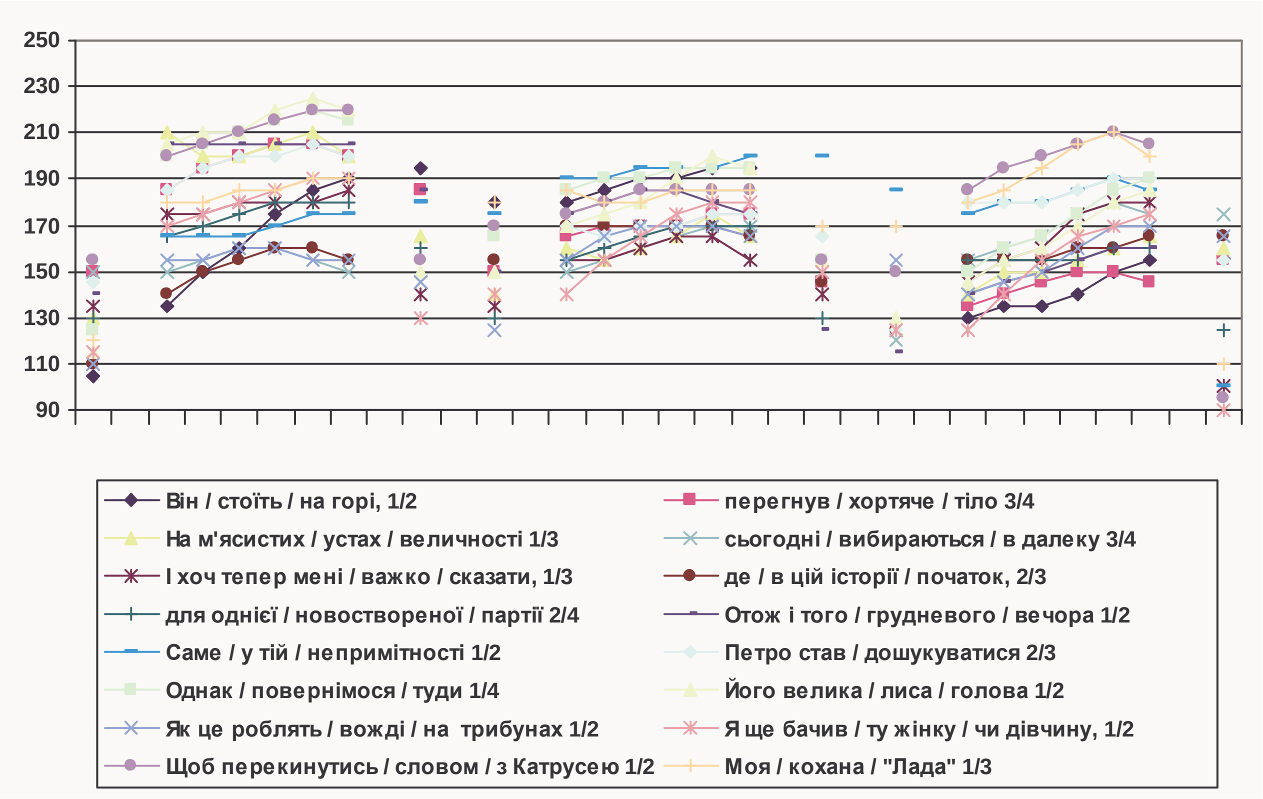

Figure 3. Non-finality intonation groups composed of three accent groups (speaker S.)

In Figure 4 different intonation contours of finality pronounced by three different speakers are shown. These intonation groups are composed of 5 accent groups.

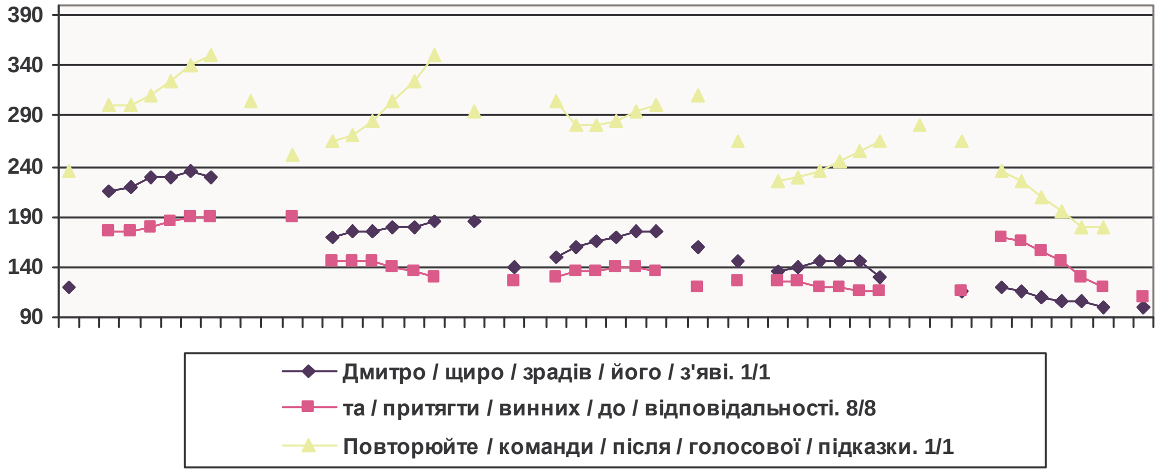

Figure 4. Finality intonation group composed of 5 accent groups (speaker S., D., and V.)

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.

5. Target specification

The trained linguistic processor gives as output the phonetic-prosodic transcription which is used later by unit selection module.

To obtain the phonetic-prosodic transcription of the input text the following procedures are involved:

- text normalization;
- segmentation into intonation groups;
- 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.

6. Unit selection

The unit selection algorithm 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 unit selection algorithm is based on phonetic and prosodic criteria. Now the main criterion of unit selection algorithm 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 voiced and non-voiced units.

For voiced 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-voiced 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.

In Figure 5 a fragment of a unit selection specification is shown. 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.

```
1052   #-s'-o    0  124
1053   s'-o-h    (1  8.0)  (2  8.0)  (4  7.9)  (5  7.9)  (6  7.9)  (8  7.8)  (9  7.8)  (10  7.8)  (12  7.7)
1054   o-h-O     (1  7.7)  (2  7.6)  (2  7.6)  (3  7.5)  (3  7.4)  (4  7.3)  (4  7.3)  (5  7.2)  (5  7.1)
1055   h-O-d'    (1  7.1)  (2  7.0)  (3  6.8)  (4  6.7)  (5  6.6)  (6  6.5)  (7  6.4)  (8  6.3)  (10  6.3)  (11  6.4)  (12  6.4)  (13  6.4)
                 (14  6.4)  (15  6.9)
1056   O-d'-n'   (1  7.1)  (2  7.2)  (3  7.2)  (4  7.2)  (5  7.3)  (5  7.3)  (6  7.3)  (7  7.4)  (8  7.4)  (9  7.4)
8378   d'-n'-i   (1  7.5)  (2  7.5)  (3  7.5)  (3  7.5)  (4  7.6)  (5  7.6)  (6  7.6)  (6  7.7)
9188   n'-i-z'   (1  7.7)  (2  7.6)  (3  7.6)  (4  7.5)  (5  7.4)  (5  7.3)  (6  7.3)  (7  7.2)  (8  7.1)
9092   y-b'-I    (1  7.1)  (2  7.1)  (3  7.1)  (3  7.1)  (4  7.0)  (5  7.0)  (6  7.0)  (7  6.9)  (8  6.9)  (8  6.9)  (9  6.8)  (10  6.8)  (11  6.8)
                 (12  6.8)  (12  6.7)  (13  6.7)
4255   b'-I-l'   (1  6.7)  (2  6.5)  (3  6.4)  (3  6.3)  (4  6.2)  (5  6.1)  (6  6.1)  (6  6.0)  (7  6.0)  (8  6.0)  (9  6.0)  (9  6.1)  (10  6.4)
771    I-l'-sh   (1  6.7)  (2  6.7)  (3  6.7)  (3  6.7)  (4  6.7)  (5  6.7)  (6  6.7)  (6  6.7)  (7  6.7)
4894   l'-sh-e   0  151
4348   sh-e-z    (1  6.9)  (3  6.9)  (4  6.9)  (6  6.9)  (7  6.9)  (9  6.9)  (10  6.9)  (12  6.9)
```

Figure 5. Unit selection module output for word combination “сьогодні більше”.

7. Unit concatenation

Unit concatenation is done by the acoustic processor in time domain. The 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 [1].

8. 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.

Large speech databases are able to retain the characteristics of the donor speaker, and the described concatenative synthesis technique is able to reflect the speaker specific characteristics in the synthesized speech.

9. References

[1] 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.

[2] 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.