The IVO Software Blizzard 2007 Entry: Improving Ivona Speech Synthesis System

Michal Kaszczuk and Lukasz Osowski

mkaszczuk@ivosoftware.com, losowski@ivosoftware.com

IVO Software Sp. z o. o.
al. Zwycięstwa 96/98, 81-451 Gdynia, Poland
http://www.ivosoftware.com

Abstract

In this paper we wish to describe special version of Ivona Speech Synthesis System with US English voice developed in IVO Software for The Blizzard Challenge 2007. Current system is based on improved speech synthesis technique originally developed for the previous challenge - The Blizzard Challenge 2006. An evaluation made by different Blizzard listeners groups, which gave the highest Mean Opinion Score to Ivona shows us, that nowadays Ivona is one of the top of available Text To Speech solutions.

Hence we show a basic overview of the Ivona Speech Synthesis System, methodology and problems which we experienced during building US English voice from the ATR database prepared for Blizzard Challenge 2007. We also show a short analysis of Blizzard Challenge 2007 results and future plans of development for Ivona Speech Synthesis System.

Index Terms: speech synthesis, Ivona Speech Synthesis System, Blizzard Challenge.

1. Introduction

The main goal of taking part in Blizzard Challenge was to compare our technology used in Ivona Speech Synthesis System with other best available solutions and their progress made during last year. When building Ivona we focused on getting best possible quality. Our customers use synthesized speech in sophisticated solutions, because of that we decided not to use any vocoding techniques and focus on full database.

The Ivona Speech Synthesis System was developed in IVO Software, Poland. In 2006 Ivona was recognized as leading text to speech technology in Mean Opinion Score given by Blizzard Challenge 2006 listeners groups. In current system we had introduced a novel approach for automated pitchmark correction and a lot of minor improvements.

Nowadays Ivona Speech Synthesis System is very well prepared commercial solution, one could say, that it is technologically mature.

On the web page http://ivona.ivosoftware.com we published an on-line version of commercial Ivona Speech Synthesis System.

Ivona Speech Synthesis System has the following features:

- Semi-automatic voice building environment.
- Very natural sounding speech.
- Fast speech production and advanced streaming technology which allows using the system in large and sophisticated implementations.

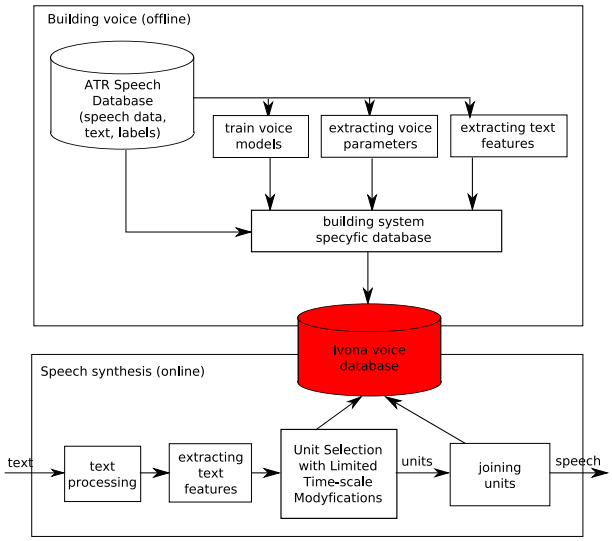

Figure 1: *An overview of the Ivona Speech Synthesis System.*

- Support for multiple languages which can be easily built and added.

The first non-Polish voice for Ivona Speech Synthesis System was US English voice developed for Blizzard Challenge 2006. Currently Ivona provides high quality US English female voice Jennifer and Romanian voice Carmen as well. The voice presented at The Blizzard Challenge 2007 is based on recorded in ATR Institute sentences. We have build this voice from scratch in less than two weeks.

We are very glad of the fact that almost all groups of Blizzard's listeners evaluated Ivona's US English voice with the highest note.

2. An Overview of the Ivona Speech Synthesis System

Ivona works very similar to common known unit selection speech synthesis scheme.

This scheme consists of two phases:

Voice building is an offline phase. During this one we extract voice parameters and text features. Then we use them to train voice dependent model such as stress and duration models. The final result is a speech database and models.

They are used during Ivona's Speech Synthesis process to generate speech.
This process detailed is described in section 3.

Speech synthesis is an online phase. In the passage of this stage Ivona produces speech from input text. There are several algorithms responsible for:

1. text analysis and processing,
2. extracting text features for model cost function,
3. finding F_0 and duration contour,
4. selecting units (polyphones[1]) from a speech database using model and concatenation cost functions,
5. modifying selected units according to contours,
6. concatenating units into speech signal.

We introduced in Ivona Unit Selection algorithm with Limited Time-scale Modifications (*USLTM*).
USLTM is based on cost function, which is responsible for selecting best units from database next used to concatenation. It also provides time-scale modifications to maintain control over the selected units' duration. The cost function consists of two elements: namely, model cost function and concatenation cost function.

$$cost(u) = model\_cost(u) + concatenation\_cost(u) \quad (1)$$

where u stands for a speech database unit. Model cost function works in phoneme domain and uses a vector of ≈ 40 features extracted from text such as phonetic context, stress and accent or phone position in hierarchy of utterance, phrase, word and syllable.
Second function - concatenation cost function is responsible for minimizing differences between concatenated units in sound "quality" domain. For this purpose concatenation cost function uses following candidate unit sound parameters:

- F_0,
- power,
- voiceness (voices/unvoiced decision),
- length,
- cepstrum coefficients normalized to 16-point curve interpolated using spline algorithm,

For unit database search a very effective Dynamic Programming algorithm is used, which makes full search of all possible candidate units combinations in near realtime.
However, serious differences between selected units and duration model sometimes occurs. To handle this we used time-scale modification algorithm as a part of USLTM. This method works in time domain, in pitch synchronous way and modifies speech without any contaminations.
Selected and modified units are then concatenated in time domain in pitch synchronous way. Overlap and Add (OLA) method is used.

[1]Polyphone stands for a group of adjecent phones.

3. Building US English voice for Blizzard Challenge 2007

US English voice for Ivona Speech Synthesis System was based and developed on speech database released by ATR Institute. This is an about five hour long recording of American English voice talent which provides about 6500 sentences. Quality of this recording is very important for final quality of the overall speech synthesis system. In this section we show the methodology of building voice. During this process we experienced some problems with database. We decided to describe few of them and we hope that it would be useful in next editions of Blizzard Challenge.
A main goal of Ivona Speech Synthesis System is to achieve the best quality of speech, so we decided to focus on full set of sentences available in ATR database.

3.1. Building methodology

US English voice has few modules similar to Polish such as text processing module. So it was easier to implement following steps:

Prepare text data using text processing and letter-to-sound rules. To do that for the Blizzard Challenge purposes we used rules and dictionaries available in Festival Speech System.

Autolabel speech recordings with pause synchronization. In this stage Sphinx autoaligner was used, but produced labels was additionally processed to resolve pause disambiguations.

Build text features vector. Feature vectors are extracted for every phone and contains over 40 miscellaneous entries.

Build voice dependent models i.e. duration model. Decision trees are trained using features extracted from text.

Prepare Ivona specific data which consists of speech units database and trained models. Units database internal structure is optimized for DP search algorithm.

Before we had started voice building process we had to solve several speech database problems which are described below. Two major problems: non-ordinary words[1] and power of recording[1] are present in current database as well.

3.2. Statistical Pitchmark Correction Method

We observed also another problem which is referred to voice characteristic used in recording of ATR database. The voice is prone to pitchmark position errors, which is critical for speech concatenation using pitch synchronous methods (i.e. PS-OLA). Advanced pitchmark labeling algorithms produces up to 10 errors per 5 seconds of each utterance (error rate is dependent of voice characteristic and utterance as well). These errors mostly occure when an algorithm locates more or less glottal closure instants in given part of speech signal then in reality. Then we can finally observe:

1. missing pitchmark in glottal closure instant,
2. multiple pitchmarks near glottal closure instant.

Having based on that observation we decided to implement simple method to correct these errors. We called it Statistical Pitchmark Correction Method.

First we need to prepare vector V (contains one value per pitchmark) using formula:

$$V_i = \frac{2 * N * \Delta t_i}{\sum_{j=i-N}^{j=j+N-1} \Delta t_j} \quad (2)$$

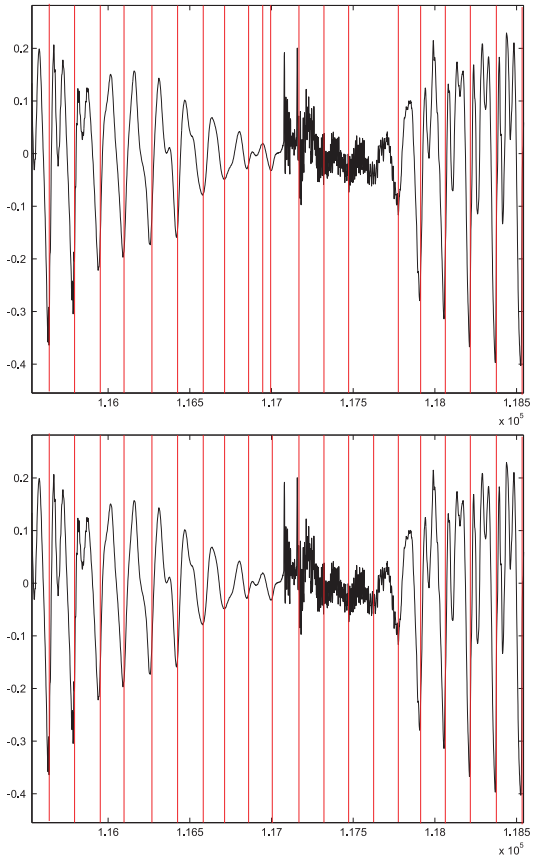

Figure 2: *Correcting pitchmarks using Statistical Correction Method.*

where:
Δt_i - pitch period duration of pitchmark i; $\Delta t_i = t_{i+1} - t_i$
N - stands for window length $2 * N$

V_i stands for pitch period duration of pitchmark i referenced to average pitch period duration (for window size $2 * N$)

Voice frequency doesn't change dramatically in regular speech, so the value V_i should change smoothly in time and oscillate near 1. Basing on it we are able to detect pitchmark problems. Value of V_i close or bigger than 2 means there probably is a missing pitchmark, V_i value close to 0 means there probably are multiplied pitchmarks.

Having used this simple method in Ivona Speech Synthesis we reduced concatenation errors and gained "smoother" sounding speech.

4. Results of Blizzard Challenge 2007

The Blizzard Challenge 2007 shows that introduced improvements let us achieve our goals. Ivona Speech Synthesis System (P) gained highest Overall Mean Opinion Score (MOS) for full dataset results (voice A).

It can be seen in table 1, that almost all listeners groups evaluated speech synthesis techniqe used in Ivona in the top of Mean Opinion Score. We are glad of the MOS results, especially of the MOS results in Volunteers group (R), which represents target group of Ivona Speech Synthesis System users. We suspect, that all listeners groups are very sensitive for

Table 1: *Mean Opinion Score (MOS) for different listeners groups (K - paid UK students, R - volunteers, S - speech experts, U - paid US students), full database.*

System	**Overall**	K	R	S	U
A	3.8	3.4	3.6	4.2	3.4
B	3.0	2.7	2.9	3.1	3.1
C	3.2	3.0	2.9	3.4	2.9
D	2.6	2.2	2.3	2.9	2.1
E	3.0	3.0	2.8	3.1	2.5
F	1.5	1.6	1.4	1.4	1.7
G	1.4	1.5	1.4	1.4	1.2
H	3.2	3.0	3.0	3.4	3.1
Voice talent	4.7	4.6	4.7	4.8	4.3
J	3.4	3.1	3.5	3.4	3.5
K	3.6	3.4	3.5	3.7	3.2
L	1.3	1.3	1.4	1.3	1.1
M	3.0	2.5	2.6	3.4	2.8
N	2.7	2.2	2.7	2.9	2.3
O	2.5	3.3	3.5	3.7	3.2
Ivona	**3.9**	**3.6**	**3.8**	**4.1**	**3.7**
Q	2.5	2.3	2.4	2.5	2.4

all speech distortions and they expect very natural human-like sounding speech. The Word Error Rate result ($median = 0.14$, $mean = 0.29$) is very good, although tuning WER was very difficult for us, because we are not native English speakers. We are convinced that most WER related problems could be solved by improving speech database and recorded corpora as well. We found Similarity reports very interesing, our system was evaluated as very similar to original speaker ($median = 4$, $mean = 4.0$). Used in IVONA USLTM algorithm performs modification of original speech in very limited range, which allows to keep orginal recording characteristic. Current system uses models trained on database provided for Blizzard Challenge as well.

Table 2: *Mean Opinion Score (MOS) with reference to voice talent score for different listeners groups (K - paid UK students, R - volunteers, S - speech experts, U - paid US students)*

System	**Overall**	K	R	S	U
Ivona	0.8297	0.7826	0.8085	0.8541	0.8604

In the table 2 we introduced Mean Opinion Score (MOS) with reference to the voice talent score. This value could be named "naturalness", and lets us know how many "naturalness" each system has. In most listeners groups Ivona Speech Synthesis System gained result 0.8297, but the most satisfying is naturalness score gained in Speech Experts group - over 0.8541, which means Ivona produces natural comparable speech.

Table 3 presents Mean Opinion Score gained by current Ivona Speech Synthesis System in caparison to last year's system presented on The Blizzard Challenge 2006. All listeners groups evaluated current system much better. The most significant progress is noticeable in Speech Experts group. Probably the most important change versus previous Blizzard entry is bigger database and statistical method of pitchmark correction. We are sure that dozens of minor improvements we did are very important as well.

Table 3: *Comparison of Ivona's Mean Opinion Score (MOS) with score achieved on previous Blizzard Challenge 2006 for different listeners groups (K - paid UK students, R - volunteers, S - speech experts, U - paid US students)*

System	**Overall**	K	R	S	U
Blizzard 2007	3.9	3.6	3.8	4.1	3.7
Blizzard 2006	3.6	-	3.5	3.7	3.7

5. Conclusions

The Blizzard Challenge 2007 results prove that Unit Selection algorithm with Limited Time-scale Modifications (USLTM) technique used in Ivona is currently one of best speech synthesis solutions, especially as far as naturalness and sound quality are concerned.

We have also noticed that Ivona's speech quality grows due to increase of database size. On the other hand we believe that current database's size is near to optimal, and further increase won't be reflected in much better quality.

The Blizzard Challenge 2007 results show that Ivona Speech Synthesis System is ready for adding new languages very easy and very quickly. The US English voice prepared from ATR speech database had been built in two weeks. Thanks to that, Ivona's cost functions in unit selection algorithm are universal we didn't have to modify it during US English voice building process. They seem to be independent from language and voice.

Participation in the Blizzard Challenge was great benefit to our system, because it gives us significant insight of what direction should we choose and what modules should be improved in our furhter development plans.

5.1. Future plans

Algorithms and tools used in Ivona Speech Synthesizer are constantly being improved, however, we focus on two main directions:

1. produce speech even more natural including improvements in NLP and USLTM,
2. fully automatic system for building new voices and languages.

6. Acknowledgments

Authors would like to thank Professor Alan W Black and all the authors of the Festival Speech Synthesis System and common tools. Their work is very important for us because it lets us learn about speech synthesis in practice. Their work was the very beginning of most of our ideas.
We are honoured to be in company of such good quality systems this year in Blizzard Challenge.
Thanks a lot!

7. References

[1] Kaszczuk, M. and Osowski, L., "Evaluating Ivona Speech Synthesis System for Blizzard Challenge 2006", Blizzard Workshop, 2006, Pittsburgh, PA

[2] Kominek, J. and Black, A., "The CMU ARCTIC Speech Databases", SSW5, 2005, Pittsburgh, PA

[3] Bennet, C. L., "Large Scale Evaluation of Corpus-based Synthesisers: Results and Lessons from the Blizzard Challenge 2005", Interspeech 2005, Lisbon, Portugal

[4] Hunt, A.J and Black, A., "Unit selection in concatenative speech synthesis using a large speech database", ICASSP, 1996

[5] Kaszczuk, M., "Opis budowy i implementacja systemu syntezy mowy polskiej Piko", Technical University of Gdansk, 2003, Gdansk, Poland

[6] Osowski, L., "System syntezy mowy polskiej", Technical University of Gdansk, 2001, Gdansk, Poland

[7] Tadeusiewicz, R., "Sygnal mowy", Wydawnictwa Komunikacji i Lacznosci, 1988, Warszawa, Poland

[8] Black, A. and Tokuda, K., "The Blizzard Challenge 2005: Evaluating Corpus-Based Speech Synthesis on Common Datasets", Interspeech 2005, Lisbon, Portugal

[9] Tokuda, K., Yoshimura, T. Masuko, T., Kobayashi, T., Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis", ICASSP, 2000, Isanbul, Turkey

[10] Hunt, A. and Black, A., "Unit selection in a concatenative speech synthesis system using unit selection synthesizer", 5th ISCA Speech Synthesis Workshop, 2004, Pittsburgh, PA

[11] Black, A. and Lenzo, K., "Optimal data selection for unit selection synthesis", 4th ISCA Speech Synthesis Workshop, 2001, Scotland