Evaluating Ivona Speech Synthesis System for Blizzard Challenge 2006

Michal Kaszczuk and Lukasz Osowski

mkaszczuk@ivo.pl, losowski@ivo.pl

IVO Software Sp. z o. o.
al. Zwyciestwa 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 2006. An evaluation made by Speech Experts group, which gave the highest note to Ivona shows us, that nowadays Ivona is in the top of Text To Speech solutions available. 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 database prepared for Blizzard Challenge 2006. We also show a short analysis of Blizzard Challenge 2006 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 starting in Blizzard Challenge was to compare our technology used in Ivona Speech Synthesis System with other best available solutions. Building Ivona we focused on getting best possible quality. Our customers use synthesized speech in sophisticed 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. Our first product - a polish speech synthesizer named ”Spiker” based on concatenation of diphones and RELPC coding was created in 2001. This product has a lot of advantages: namely, it produces a very clear and fluent speech, it works very fast and meet the requirements of most portable devices such as mobile phones and PDAs. In the years 2001 - 2004 Spiker was the best sold polish speech synthesizer. But the technology used in Spiker wasn't good enough, so we decided to develop brand new speech synthesis system which could produce a very natural sounding speech. The first commercial version of a new system - Ivona Speech Synthesis System was finished in the half of 2005. Since then technology and the overall system have been continuously improved to achieve the best results.

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

Ivona Speech Synthesis System has the following features:

- Very natural sounding speech.
- Fast speech production and advanced streaming technology which allows using the system in large and sophisticated installations.
- Support for multiple languages which can be easily build and added.

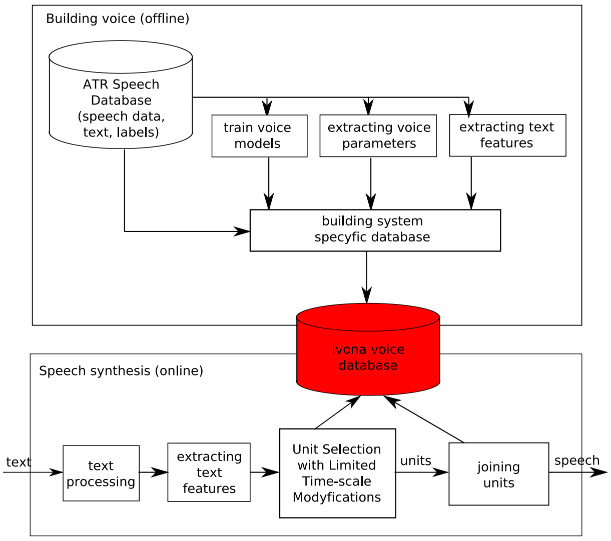
Building voice (offline)
ATR Speech Database (speech data, text, labels)
train voice models
extracting voice parameters
extracting text features
building system specyfic database
Ivona voice database
Speech synthesis (online)
text
text processing
extracting text features
Unit Selection with Limited Time-scale Modyfications
units
joining units
speech

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

The first non-polish voice for Ivona Speech Synthesis System is US English voice developed for Blizzard Challenge 2006. It is based on recorded in ATR Institute sentences. We build this voice from scratch in two weeks.

We are very glad of the fact that a group of Speech Experts 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 processing,
2. extracting text features for model cost function,
3. finding F_0 and duration contour,
4. selecting units (poliphones) 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.

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

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.

3. Building US English voice for Blizzard Challenge 2006

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 4273 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.

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

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

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

Before we started voice building process we had had to solve several speech database problems which are described below.

3.2. Non-ordinary words

Table 1: *Sentences in ATR database.*

set	no. sentences	no. non-regular sentences
ARCA	542	42
ARCB	538	34
BTECH	2129	286
NEWSPAPER	1049	367

In recorded speech there was a lot of non-ordinary words not easily found in American English language which constituted a problem for us, for example *"Is that Hiroshi Suzuki, Masao Suzuki or Taro Suzuki?"* (recording BTEC 00220). We realize, that Japanese pronunciation is very similar to English, but phonetic context and the way American voice talent read that is not natural. In table 1 we show a comparison of number of sentences in selected recording sets against a number of sentences containing non-ordinary words. Summing up you should notice that over 17% of all recorded sentences contain practically unusable words. Based on above some part of recorded sentences were dropped, especially those with the highest density of non-ordinary words.

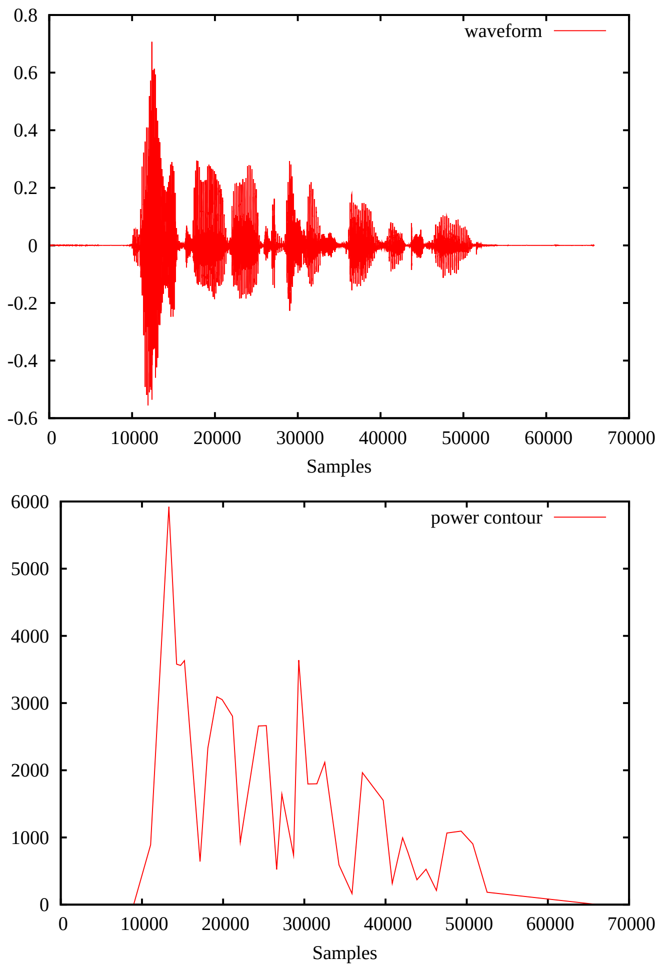

Figure 2: *A waveform and a power contour of recording ARCA 00014.*

3.3. Power

The biggest problem with ATR speech database was a power contour of recorded sentences. The sentences differ between each other in power. There is also a lot of sentences with power contour problems within (figure 2). Using this sentences without solving power contour problems may cause strange phenomena in produced speech such as some parts of speech could be lauder the others.

To solve the power contour problem we made few things. First we found normalization factors which allows to minimize differences between phrases. Then we implemented additional condition within USLTM algorithm. This approach has online character and allows us to determine what units match power domain. As far as our opinion is concerned this method gives great effects because input sentences has it's natural power contours. However, tests results show us, power condition used in unit selection algorithm determines denying of $\approx$ 30% units which are well matched in other conditions. The same situation is when speech database is $\approx$ 30% smaller but recorded without power contour problems.

Presented problems have a big influence on final database size used in speech synthesis process. For Ivona's Polish voice Jacek we achieve similar quality with database which contains about 1500 well selected and recorded sentences.

4. Results of Blizzard Challenge 2006

The Blizzard Challenge 2006 shows that we achieved our goals. Our system (K) is the best in Mean Opinion Score (MOS) provided by Speech Experts and Undergraduates for full database results. Speech Experts evaluated Ivona as clear winner in Mean Opinion Score. There was a huge distance between speech produced by Ivona and those produced by other systems in evaluation made by Speech Experts group. We suspect, that Speech Experts are very sensitive for all possible mistakes and errors in synthsiesed speech. Also they do not expect vocoded speech, which cannot be used in sophisticed applications where a sound quality is of much importance.

We are glad of Word Error Rate (WER) results, i.e. Undergraduates choose Ivona as the second best system.

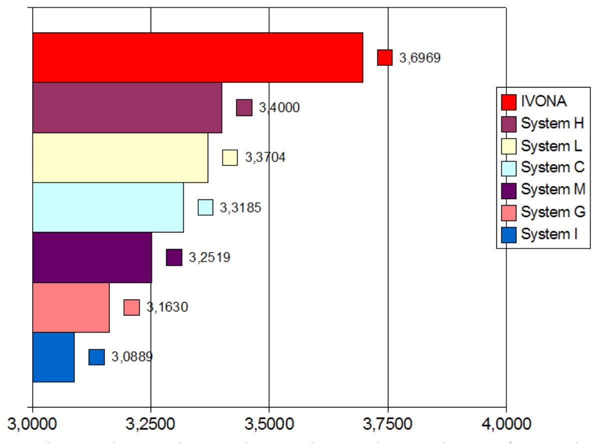

Figure 3: *Mean Opinion Score (MOS) of 7 best systems in Speech Experts group.*

On the table 3 we introduced Mean Opinion Score (MOS) with reference to voice talent score. Those values could be named "naturalness", and lets us know how many "naturalness" each system has. In every listeners group Ivona Speech Synthesis System gained result $\approx$ 80%, which means that speech produced by our system is near natural human speech.

5. Conclusions

The Blizzard Challenge 2006 results proves that Unit Selection algorithm with Limited Time-scale Modifications (USLTM) techinque used in Ivona is currently one of best speech synthesis solutions, especially in naturalness and sound quality.

The Blizzard Challenge 2006 results show that Ivona Speech Synthesis System is ready for adding new languages quickly. The US English voice prepared from ATR speech database was build in two weeks. Thanks to Ivona's cost functions in unit selection algorithm being universal we didn't have to modify it during US English voice building process. They seem to be independent from language and voice.

Table 2: *Mean Opinion Score (MOS) for different listeners groups (S - Speech Experts, U - Undergraduates, R - Random), full database.*

System	S	U	R
A	2.9259	2.9405	3.0000
B	2.4667	2.5595	2.4576
C	3.3185	3.7262	3.5141
D	2.9481	2.8810	2.8023
E	1.3926	1.6190	1.5706
F	2.9481	2.7381	2.6328
G	3.1630	3.2381	3.1243
H	3.4000	3.5357	3.1695
I	3.0889	3.3690	3.0169
J	2.0000	2.0119	1.9435
Ivona	**3.6963**	**3.7381**	**3.4576**
L	3.3704	3.3810	3.2034
M	3.2519	3.4405	3.2203
N	2.5185	2.5357	2.4407
Voice talent	4.6593	4.4405	4.5141

Table 3: *Mean Opinion Score (MOS) with reference to voice talent score for different listeners groups (S - Speech Experts, U - Undergraduates, R - Random), full database.*

System	S	U	R
The best of other systems	0.7297	0.8391	0.7784
Ivona	**0.7933**	**0.8418**	**0.7659**

5.1. Future plans

Algorithms and tools used in Ivona Speech Synthesiser 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 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 to learn about speech synthesis in practice. Their work was the very beginning of most of our ideas.
Thanks a lot!

7. References

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

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

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

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

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

[6] Tadeusiewicz, R., ”Sygnal mowy”, Wydawnictwa Komunikacji i Lacznosci, 1988, Warszawa, Poland

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

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

[9] 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

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