The IVO Software Blizzard Challenge 2009 Entry: Improving IVONA Text-To-Speech

Michal Kaszczuk[1], *Lukasz Osowski*[1]

[1]IVO Software Sp. z o. o.
al. Zwyciestwa 96/98, 81-451 Gdynia, Poland
http://www.ivosoftware.com
mkaszczuk@ivona.com, losowski@ivona.com

Abstract

This paper describes a special version of IVONA Text-To-Speech for a GB English voice designed and developed by IVO Software for The Blizzard Challenge 2009. The architecture of this system is based on an improved IVONA Text-To-Speech originally developed for previous challenges - Blizzard Challenge 2006[1] and Blizzard Challenge 2007[2].
This year we decided to build two GB English systems (using the full database and the arctic subset) and complete four challenge tasks EH1, EH2, ES2 and ES3. The system used for completing tasks E21 and ES3 as well as for task EH1 was built on the full 'roger' database.
Hence we show a basic overview of the IVONA Text-To-Speech architecture. Then we focus on methodology and problems which we experienced during development of our GB English voice from the 'roger' database provided by CSTR[1]. We also present a short analysis of the Blizzard Challenge 2009 results and future plans for development of IVONA Text-To-Speech.
Index Terms: IVONA, IVO Software, Speech Synthesis, Blizzard Challenge.

1. Introduction

The main goal of taking part in the Blizzard Challenge was to compare speech synthesis used in IVONA Text-To-Speech with other available solutions and their progress made the during past two years. While building the IVONA system for the challenge we focused on gaining the best possible quality and naturalness of speech out of a voice built in a semi-automatic process.
We decided to use no vocoding techniques and focus on gaining the best results from a concatenative, unit selection approach while taking advantage of the full provided speech data.

The IVONA system we built for Blizzard Challenge 2009 was developed in less than two weeks using Rapid Voice Development, our in-house semi-automatic voice development toolset. We have no native English speakers in our team.

IVONA Text-To-Speech is widely used in commercial solutions. On http://ivona.com we published an on-line version of our commercially available IVONA Text-To-Speech voices. IVONA currently provides several high-quality voices in US English, Polish and Romanian. In a few weeks[2] we are going to release two new IVONA GB English voices based on this year's Blizzard Challenge experience. More languages and voices will be available soon.

[1]Centre for Speech Technology Research, University of Edinburgh
[2]September/October 2009

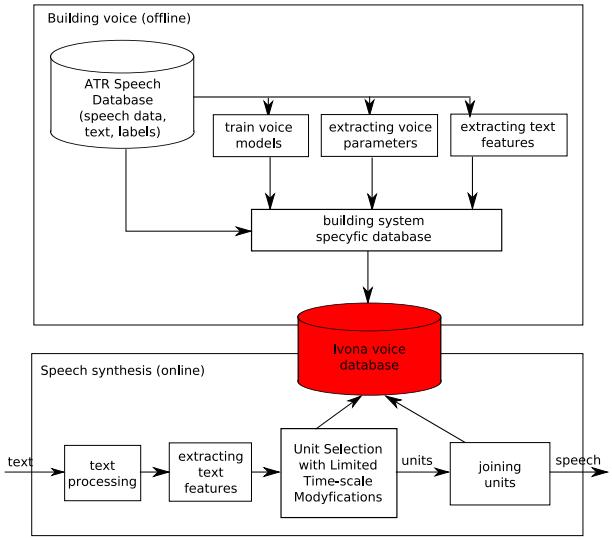

Figure 1: *An overview of the IVONA Text-To-Speech System.*

2. An Overview of the IVONA Text-To-Speech

IVONA is a concatenative system and works very similarly to commonly known unit selection speech synthesis scheme.

The diphone is the system's base sub-word unit, nevertheless the system is designed to be able to fall back to half-phones when necessary. This ensures that the system can synthesise speech even if some of the diphones appearing in text are missing from the database (or extremely do not match in the terms of audio signal quality).

The voice building is an off-line phase and is realized using the Rapid Voice Development semi-automatic process.
Using Rapid Voice Development we perform automatic segmentation of recordings basing on text prompts. We extract phonetic and linguistic features from the text prompts and audio parameters from speech recordings. The collected data is used in automatic training of voice-dependent models such as prosody prediction model.

In IVONA we implemented the Unit Selection algorithm with Limited Time-scale Modifications. USLTM is based on cost function, which is responsible for selecting the best speech units from large speech unit database. The speech units chosen from the database are next used during concatenation and production of waveform. USLTM also performs some time-scale

modifications to maintain control over the prosody parameters (duration, pitch) and power of the selected speech units.
The cost function consists of two elements: model cost function and concatenation cost function.

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

where u stands for the database speech.
The model cost function works in phoneme domain and uses a vector of ≈ 50 (language dependent) phoneme features extracted from text such as phonetic context, stress and accent, phone position in hierarchy of syllable, word, sentence and finally whole utterance.
The concatenation cost function is responsible for finding such speech units in the database which provide best match in terms of sound concatenation quality. For this purpose, the function employs various sound parameters such as F_0, power, voiceness (voices/unvoiced decision) and normalized spectrum coefficients.

An effective dynamic programming algorithm is used for unit database search, which performs a full search of all possible combinations of candidate units. However, sometimes serious differences between selected units and prosody model occur. To decrease differences we use time-scale modification algorithms as a part of USLTM. This method works in time domain in a pitch synchronous way. The system modifies speech units in very limited range in order to prevent distortions or serious deformations of original audio signal.
Selected and prepared (modified) speech units are concatenated in time domain in pitch synchronous way using the Overlap and Add (OLA) joining method.

3. Building a voice from the 'roger' database

The GB English voice for The Blizzard Challenge 2009 was built using the provided 15 hours long 'roger' database. This voice is our first GB English voice as we had no previous experience with GB English accented speech data. We decided to use Rapid Voice Development in order to build a high-quality voice very fast from the 'roger' database.

The 'roger' database consists of the following sections: carroll, unilex, address, spelling, arctic, emphasis, the herald and news (9512 sentences in total).
The quality of the database is crucial for the final speech quality generated by the system therefore a checking process of the provided data was an essential prerequisite. We decided to use sentences from all database sections however we removed 1089 sentences ($\approx 11\%$) from the database in a manual pruning process. Sentences encountering the following issues have been removed:

1. audio quality significantly differing from the rest of the database,
2. over-emphasised reading style (mostly from carroll section),
3. foreign words with ambiguous pronunciation,
4. recording mistakes of the voice talent such as laughs, noises or background talks.

We did not use the provided festival utterance files.
The phonetic translation of text prompts was required in order to perform automatic segmentation of the speech data. For this purpose we had to implement text analysis, disambiguation of homographs (pronunciation variants) and many others GB English specific linguistic issues. We utilized Unisyn-1.3 lexicon (provided with data) with our own lexicon addenda.

The database was recorded using text prompts containing a variety of *emphatic* words[8, 9] [3]: "Let this be a lesson to you, NEVER to lose YOUR temper!". In order to take advantage of this database feature we tagged phonemes deriving from emphatic words in the built speech database. USLTM was modified to prioritize emphatic words diphones during selection when such words appear in the input text or omit them otherwise.

Another improvement of our system this year is utilizing last word emphasis of sentences ending in an exclamation mark: "absurd: absurd: absurd. absurd? absurd!". This kind of word emphasis can be found in over 4500 sentences of the provided database.

4. Results

Several aspects of synthesized speech were in the focus of the 2009 Blizzard Challenge. From our system's point of view the most important is naturalness and similarity to original speaker evaluated in task EH1 (combined with ES1) and task ES3.

The letter *K* in the following results corresponds to the IVONA Blizzard Challenge 2009 participating system.

4.1. Tasks EH1 (full database)

For task EH1 we decided to focus on results gained on "News" listening tests section taking into consideration the fact that this is the only section which provides unambiguous comparing results of all participating systems in the same test domain[4].

As shown in figure 2 our system obtained one of the highest Mean Opinion Score of 4.0 while the voice talent's result was 4.9. The other highest MOS is 4.2 gained by the S system. The Pairwise Wilcoxon rank shows no significant difference in naturalness between our (K) system and the S system (table 2).
As can be noticed in table 1 the next highest MOS for naturalness is 3.5 gained by the I system which significantly differs from our (K) system result (table 2).

Figure 3 shows the results of similarity to original speaker of the systems. Our system gained the second highest Mean Opinion Score - 3.7 while the voice talent's result was 4.9. The highest similarity score was 3.8 gained by the S system. According to the Pairwise Wilcoxon test there was no significant difference in similarity between our system (K) and the S system (table 2). The next highest similarity MOS is 3.2 gained by the I system (table 1) which significantly differs from the result of our system (K).
This shows a typical correlation between in MOS for naturalness and similarity which can be noticed in unit selection speech systems, built on real speaker recordings. The participating systems which obtained highest MOS for naturalness gained the highest MOS for similarity to original speaker as well.

4.2. Task ES2 (full database, transmission through telephone line)

A very interesting section in this year's challenge was task ES2 which provides information on how evaluation results change

[3] Written in capital letters

[4] Due to a mistake in listening tests setup some samples of certain systems were evaluated by fewer listeners that it was intended. A detailed explanation has been included in the README file located in the Blizzard Challenge results provided by the challenge organizers

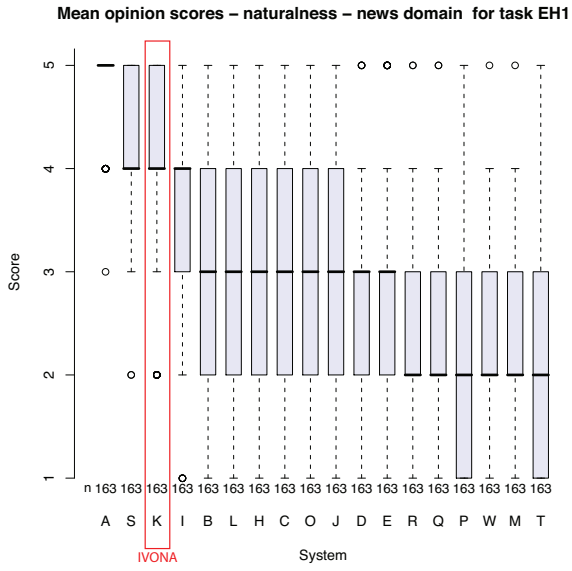

Figure 2: EH1 - Naturalness (MOS)

Table 1: *EH1 - Naturalness (MOS) for different listener groups (EU - paid participants (English-native), ER - volunteers, ES - speech experts)*

System	**Overall**	EU	ER	ES
A (Voice talent)	4.9	4.9	4.9	4.8
B	3.1	2.9	3.2	3.3
C	2.9	2.8	2.9	3.0
D	2.6	2.4	2.8	2.8
E	2.6	2.3	2.8	3.1
H	3.0	2.7	3.1	3.4
I	3.5	3.6	3.2	3.4
J	2.7	2.5	2.7	3.1
K (IVONA)	**4.0**	**3.9**	**4.0**	**4.2**
L	3.1	3.0	3.2	3.3
M	2.2	2.1	2.3	2.2
O	2.8	3.0	2.6	2.7
P	2.3	2.4	2.3	2.2
Q	2.4	2.2	2.6	2.6
R	2.5	2.5	2.4	2.6
S	4.2	4.2	4.1	4.3
T	2.1	2.0	2.3	2.1
W	2.3	2.1	2.5	2.5

while speech is transmitted through telephone line. It is significant that the results of all systems including the voice talent (system A) are lower in reference to the EH1 task results.

In this task our system (K) gained the highest Mean Opinion Score for naturalness - 3.5 (table 5) while the second best result, 3.4, was gained by the S system. The Pairwise Wilcoxon signed rank in table 6 shows that there is no significant difference in naturalness between our system (K) and the S system. The above-mentioned table also shows that all other systems differ significantly in naturalness from our (K) system and the S system.

The most interesting issue in the ES2 task results is the Word Error Rate section. As shown in table 7 our system gained the best result - the lowest error rate of 0.31. We believe this is an effect of USLTM used in IVONA which modifies prosody parameters of concatenated units in very limited range. This results in fewer speech audio signal distortions which could be emphasised while being transmitted through a telephone line.

5. Conclusions

In terms of naturalness and similarity to original speaker the only system which gained a better Mean Opinion Score than our system and at the same time its result is significantly better, according to Pairwise Wilcoxon signed rank, is the voice talent. We believe the above proves that our Unit Selection algorithm with Limited Time-scale Modifications (USLTM) used in IVONA is currently one of best speech synthesis techniques, especially as far as naturalness, similarity to original speaker and sound quality are concerned.

The participation in the Blizzard Challenge is always of great benefit to our system as it gives us a significant insight of what direction should we choose and what modules of the system should be improved. It also gives us the opportunity to test our algorithms and Rapid Voice Development scheme on non in-house speech databases (of specific design and recording methodology) and check at what extent our system depends on the quality of the database.

This year's Blizzard Challenge gave us possibility to experiment on one of the biggest database in our Text-To-Speech experience. Although the speech quality of concatenative unit selection speech synthesisers (such as IVONA) depends on unit database size, we have noticed no significant increase of the speech quality using the 'roger' (15 hours long) database.

Our last conclusion is that the Rapid Voice Development - semi-automatic voice building process is a very effective method. Making use of RVD we were able to create our system for the Blizzard Challenge in a very short period of time obtaining very good results in the listening tests.

5.1. Future plans

The algorithms and tools used in IVONA Text-To-Speech and during the voice building process (such as Rapid Voice Development) are constantly being improved. However, we focus on improving our USLTM method to generate speech more natural and more similar to original speaker.
We are also planing some improvement works in Rapid Voice Development in order to create fully automatic (or more "automatic" in "semi-automatic") process of building new voices and language models without any expertize in speech synthesis, linguistics or phonetics.

6. Acknowledgments

We 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 because it gives us the possibility to learn new things about speech synthesis in practice.
We would like to thank all Blizzard Challenge 2009 organizers for setting up such a great speech synthesis challenge and donating all required resources such as the 'roger' speech database.
We would also like to thank Remus Mois and Michal Czuczman, members of the Research and Development team of IVO Software for their the greatest effort put in our entry.
Thanks a lot!

Table 2: *EH1 - Naturalness - Significant differences between systems (Pairwise Wilcoxon signed rank - TRUE means significant difference)*

System	K
A (Voice talent)	TRUE
B	TRUE
C	TRUE
D	TRUE
E	TRUE
H	TRUE
I	TRUE
J	TRUE
L	TRUE
M	TRUE
O	TRUE
P	TRUE
Q	TRUE
R	TRUE
S	FALSE
T	TRUE
W	TRUE

7. References

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

[2] Kaszczuk, M. and Osowski, L., "The IVO Software Blizzard 2007 Entry: Improving IVONA Speech Synthesis System" Blizzard Workshop, 2007, Bonn, Germany.

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

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

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

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

[7] Karaiskos, V., King, S., Clark, R., Mayo C., "The Blizzard Challenge 2008", Blizzard Workshop, 2008, Brisbane, Australia.

[8] Strom, V., Nenkova, A., Clark, R., Vazquez-Alvarez, Y., Brenier, J., King, S., Jurafsky, D., "Modeling Prominence and Emphasis Improves Unit-Selection Synthesis".

[9] Strom, V., Clark, R., King, S., "Expressive Prosody for Unit-selection Speech Synthesis".

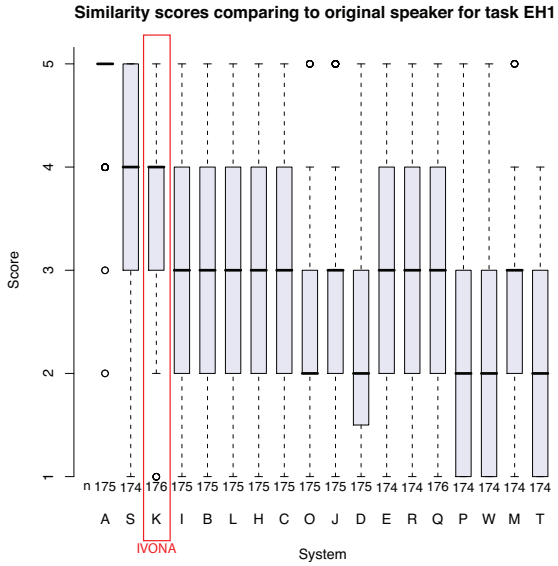

Figure 3: EH1 - Similarity to original speaker (MOS)

Table 3: *EH1 - Similarity (MOS) for different listener groups (EU - paid participants (English-native), ER - volunteers, ES - speech experts)*

System	**Overall**	EU	ER	ES
A (Voice talent)	4.9	4.9	4.8	4.9
B	3.1	2.7	2.7	3.4
C	2.9	2.9	2.9	2.9
D	2.4	2.5	2.5	2.5
E	2.9	2.9	2.9	3.2
H	3.0	2.9	2.9	3.5
I	3.2	3.2	3.2	3.4
J	2.8	2.6	2.6	3.1
K (IVONA)	**3.7**	**3.7**	**3.5**	**3.9**
L	3.0	3.1	2.9	2.8
M	2.6	2.7	2.3	2.6
O	2.5	2.6	2.4	2.6
P	2.0	2.0	1.9	2.0
Q	2.6	2.6	2.3	2.9
R	2.9	2.9	2.5	3.1
S	3.8	3.9	3.8	3.7
T	1.9	1.8	2.0	1.9
W	2.2	2.1	2.3	2.3

Table 4: *EH1 - Similarity - Significant differences between systems (Pairwise Wilcoxon signed rank - TRUE means significant difference)*

System	K
A (Voice talent)	TRUE
B	TRUE
C	TRUE
D	TRUE
E	TRUE
H	TRUE
I	TRUE
J	TRUE
L	TRUE
M	TRUE
O	TRUE
P	TRUE
Q	TRUE
R	TRUE
S	FALSE
T	TRUE
W	TRUE

Table 5: *ES2 - Naturalness (MOS) for different listener groups (EU - paid participants (English-native), ER - volunteers, ES - speech experts)*

System	**Overall**	EU	ER	ES
A (Voice talent)	4.3	4.5	4.1	4.0
B	2.3	2.3	2.4	2.2
C	2.6	2.6	2.7	2.7
D	2.6	2.5	2.8	2.6
I	2.8	3.1	2.5	2.5
K (IVONA)	**3.5**	**3.6**	**3.5**	**3.4**
L	3.0	3.0	3.1	3.0
O	2.8	2.9	2.6	2.9
P	1.9	1.9	2.1	1.9
Q	2.3	2.3	2.6	2.2
R	2.0	2.1	2.2	1.9
S	3.4	3.4	3.5	3.4
W	2.3	2.3	2.5	2.3

Table 6: *ES2 - Naturalness - Significant differences between systems (Pairwise Wilcoxon signed rank - TRUE means significant difference)*

System	K
A (Voice talent)	TRUE
B	TRUE
C	TRUE
D	TRUE
I	TRUE
L	TRUE
O	TRUE
P	TRUE
Q	TRUE
R	TRUE
S	FALSE
W	TRUE

Table 7: *ES2 - Word error rate for different listener groups (EU - paid participants (English-native), ER - volunteers, ES - speech experts)*

System	WER
A (Voice talent)	0.20
B	0.42
C	0.37
D	0.50
I	0.54
K (IVONA)	**0.31**
L	0.36
O	0.40
P	0.45
Q	0.51
R	0.55
S	0.38
W	0.39