Evaluation of synthetic speech using automatic speech recognition

Jolanta Bachan, *Adam Mickiewicz University in Poznań*
Tomasz Kuczmarski, *Adam Mickiewicz University in Poznań*
Piotr Francuzik, *Poznań Supercomputing and Networking Center*

Abstract

The present paper presents a novel method of speech synthesis evaluation by a speech recognition system. For the experiment, 4 different speech synthesis technologies were used, including diphone concatenative, statistical parametric and two unit-selection speech synthesisers. For the evaluation, the Polish speech recognition system was used. The results showed that the statistical parametric and unit-selection synthesisers produced speech which had the highest recognition accuracy. Speech synthesis based on the concatenation of diphones had the lowest recognition rate.

1. Introduction

Speech Application Programming Interface (SAPI) is rapidly becoming the most convenient solution for a wide span of electronic devices. Speech synthesisers and speech recognition systems are widely installed in personal computers, mobile phones, PDAs, cars, and intelligent houses, not to mention more advanced applications, such as assistive technologies and conversational robots.

The present paper reports on a preliminary experiment in which various speech synthesisers were evaluated by measuring the automatic recognition accuracy of their output.

In the experiment, 4 speech synthesisers were used. Two of them were developed by the authors for their M.A. and PhD theses, Automatic Close Copy Speech (ACCS) synthesis with MBROLA [1] and statistical parametric synthesis (HTS) [2]. They were compared with fully-developed systems for Polish, BOSS [3] and commercial Ivona™ [4,5]. The synthetic speech outputs of those systems were input to the Poznań Automatic Speech Recognition (ASR) system [6].

The research was carried out in order to see how synthetic speech could perform in real communicative situations where at least one of the participants is an artificial agent, for example, between a human using a speech synthesiser as an alternative or augmentative communication method and an automatic dialogue system.

2. Speech synthesisers

2.1 ACCS synthesis with MBROLA

A conventional Text to Speech (TTS) synthesis architecture has two main components: the Natural Language Processing Component (NLP) and the Digital Signal Processing Component (DSP). In Automatic Close Copy Speech (ACCS) synthesis, however, it is the annotated speech corpus that provides all the information required by the DSP component, making the NLP front-end of the TTS system redundant. The main tasks of the NLP front-end are replaced as follows [1]:

1. *Phonetisation model:* replaced by a phoneme inventory for phonemically annotated corpora.
2. *Duration model:* from the time-stamps of the annotation.
3. *Pitch model:* pitch extraction algorithm over the given label time domains.

For the ACCS synthesis, the MBROLA diphone synthesis was chosen [7]. The input to the ACCS synthesis is a pair of speech signal files and a time-aligned phonemic annotation. The phonemes are first validated (checked whether the phonems are present in the synthetic voice). Next, durations and pitch are extracted. Finally, phoneme labels, durations and pitch positions and pitch values are integrated into the synthesiser interface format (MBROLA PHO format).

The ACCS synthesis was developed as a part of Bachan's M.A. thesis. For the synthesis, two MBROLA voices were used: the previously available PL1 female voice [8] and a PL2 male MBROLA voice which was developed for implementation in a demo dialogue system in the Ph.D. thesis [9].

2.2 Statistical parametric (HTS)

Statistical parametric speech synthesis based on Hidden Markov Models (HMM) is a fairly new approach rapidly growing in popularity. In HMM-based Speech Synthesis System (HTS) [hts 2.0], spectral parameters, excitation and duration of speech segments are simultaneously modelled by context-dependent HMMs. Speech signal is later generated from the models themselves [10].

This technology allows building small footprint, high quality synthesisers from small corpora. An additional advantage is a relatively easy manipulation of voice parameters.

The synthesiser used in the current research was primarily built as a part of an M.A. thesis, using a speech corpus designed for the Polish BOSS. The Polish BOSS synthesiser is also used as a text analysis tool for the HTS, which itself does not include one [2].

This voice is currently further developed as part of a Ph.D. dissertation on F_0 modelling.

2.3 BOSS

Bonn Open Synthesis System (BOSS) [11,12] is a general open-source framework for non-uniform unit selection synthesis. BOSS architecture follows the client-server paradigm. The client is a program that can run remotely (on any system), sending input in the XML format to the server and receiving the speech signal. BOSS performs the synthesis using a two-stage algorithm: first, the candidate selection is performed, where units of highest possible level (word, syllable or phoneme) are selected from the corpus. In the second stage, the final unit selection is made based on candidates' contextual, segmental, suprasegmental features and their joint adequacy represented by cost functions.

For the Polish BOSS, the speech material (4 hours) was read by a professional radio speaker during several recording sessions, and later annotated on word, syllable and phone levels by trained phoneticians, with suprasegmental information about prosody of utterances [3].

2.4 Ivona™

Ivona™ [4,5] is a commercial speech synthesis system developed and maintained by IVO Software™, Poland. It uses a non-uniform unit selection technology. In this approach, a vector of ca. 50 features is first extracted from a preprocessed text input. On this basis, F_0 and duration contours are generated. Next, *polyphones* are selected from a large speech database according to the model and concatenation cost functions. After applying limited time-scale modifications to selected units' pitch, duration and power, they are concatenated using Pitch Synchronous Overlap and Add (PSOLA) algorithm.

Since 2006 Ivona™ has won the Blizzard Challenge, an international speech synthesis evaluation event, several times and is currently considered to be one of the best available TTS systems.

3. ASR system

The Automatic Speech Recognition (ASR) system has been developed at Poznań Supercomputing and Networking Center (PSNC) and its aim is to serve as a Polish speech dictation system (or Large Vocabulary Continuous Speech Recognition, LVCSR) for a specific language domain [6]. The system is dedicated for the Judicature, the Police, the Border Guard and other public security services. The system is designed to assist the user in writing notes, protocols, formal texts, legal documents, police reports as well as other tasks connected with the domain.

The acoustic models for the ASR system have been trained on the Jurisdict database which contains speech material of over 2000 speakers from all parts of Poland. A recording session for one speaker provides around 40min of speech and is structured in the following manner [13]:

1. Type A – (semi-)spontaneous speech: elicited dictation of short descriptions, isolated phrases, numbers or letter sequences.
2. Type B – read speech: phonetic coverage and syntactically controlled structures. These sentences were created for research purposes to provide triphone and diphone coverage and to cover the most important syntactic structures.
3. Type C – read speech: semantically controlled structures of utterances containing specialised vocabulary, legal and police texts, application words.

The system may operate on different levels of accuracy/speed preset – the lower the level, the lower the accuracy, but the faster the time of recognition. In the present study, results on 7 levels are discussed.

Additionally, the ASR system provides a speaker adaptation procedure which renders the system speaker-dependent and improves the results of speech recognition. The system accuracy tends to saturate at ca. 91% in case of non-adapted acoustic models, whereas for adapted-speaker models the results are around 93% with ever-increasing recognition time [6].

4. Speech material

An anonymised police report[1] was chosen as a text to be synthesised. This choice was justified by the particular domain of the ASR system. The report has a typical structure and consists of 277 words.

[1] All the names in the police report are fictional and should not be linked to real people.

Beginning with general information about the suspect, and followed by his confession of the guilt and explanations, it ends with a final statement of submitting to the punishment.

The report was read and recorded by an expert user of the ASR system, Piotr. In order to provide better control and easier data analysis at all stages of the experiment, the report was divided into 44 small pieces at the end of the sentences or at pauses in the speech signal.

The original human speech was automatically annotated on the phone-level using SALIAN [14] and checked manually by a phonetician. These recordings and their annotations were input to the ACCS system and synthesised using PL1 [8] and PL2 [9] MBROLA voices. The MBROLA interface format (the MBROLA PHO file) is structured into a list of tuples of phoneme label, duration in milliseconds, and an optional series of pairs of pitch position in percent of the segment duration and F_0 value in Hz. Because the original recordings were of male voice with male F_0 values and PL1 voice is female, the pitch values extracted from the human voice were doubled to make them resemble female values, i.e. original_male_F_0 * 2 = female_F_0 input to MBROLA PHO file.

At the next stage, the report was synthesised in the HTS and BOSS systems, both built on the same database of speech – the Polish BOSS, and additionally the text was synthesised in Ivona™ using the male voice of Jan.

Altogether, the speech material consisted of one police report divided into 44 pieces, synthesised with:

1. ACCS with Mbrola – PL1 female voice
2. ACCS with Mbrola – PL2 male voice
3. BOSS – male BOSS Polish voice
4. HTS – male BOSS Polish voice
5. Ivona™ - male voice, Jan

Additionally, the speech material included the original recording of male human speech – Piotr.

5. Speech evaluation and recognition experiment

The synthetic speech materials, along with the recordings of the human speech were input to the ASR system and processed on 7 levels of system's accuracy/speed presets. Level 1 shows the lowest accuracy of recognition, but works fastest. On the other hand, Level 7 shows the best recognition results, but the recognition time is longer.

The workflow of the experiment is presented in Fig. 1. 4 synthetic voices were used (drawn in white cylinders): Jan, PL1, PL2 and Polish BOSS. PL2 voice (diphone database) is based on the natural voice of Mariusz which is included in the Jurisdict database. In the experiment the recording of natural Piotr's voice was used – Piotr was also recorded earlier and is one of the 2000 speakers who gave their voices for the Jurisdict database. (Natural voices are drawn in gray cylinders.) Although the police report read by Piotr did not appear in any recording sessions included in the Jurisdict database, the fact that the ASR acoustic models were trained also on Piotr's voice might improve the speech recognition. The connection from Piotr's voice to the original recording of the police report is marked by the dotted arrow. Piotr's recording was also used for the ACCS synthesis which is marked by the dashed arrow.

The generated synthetic speech was evaluated by the ASR system. The evaluation part in Fig. 1 is marked in hatching and includes recognition of the original human speech. The ASR outputs text data whose analysis is discussed below.

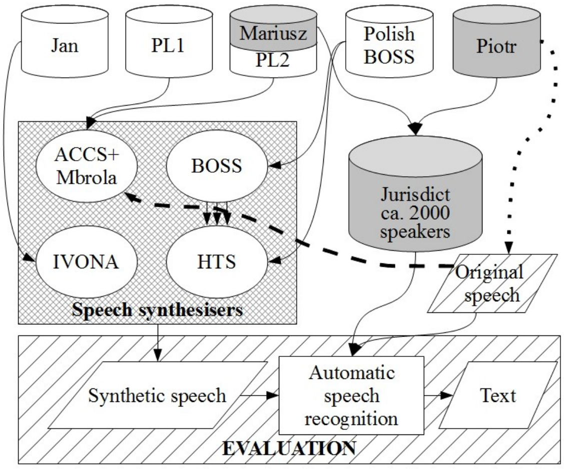

Fig. 1: The workflow of the experiment. The dotted arrow marks the connection between Piotr's voice to the original recording; the dashed arrow shows the connection of using the original recording (phones and their durations) for the ACCS synthesis with Mbrola.

6. Results

6.1 General statistics

Tab. 1 shows the results of speech recognition on 7 levels of accuracy/speed preset for 5 synthetic voices and the model human voice. On all the levels, the human speech was recognised best. Among the synthetic voices, the HTS synthesis received the best accuracy scores on almost all levels, but the real time factor was the best for the BOSS system. (The best results are marked in green bold; the worst results are in red bold.) The PL1 MBROLA voice had the worst results of accuracy recognition. Ivona™, on the other hand, had the longest recognition time.

The comparison of the speech recognition results on the lowest Level 1 and the highest Level 7 are presented in Tab. 2. The table shows summary of correct word recognition, substitutions, deletions, insertions and the total error rate. The last column presents the percentage of sentences with errors, where one wrong unit (e.g. a letter) counts the sentence as incorrectly recognised.

The human voice again received the best scores, however on Level 7 it had higher percentage of word insertions than the HTS and the BOSS syntheses.

Comparing the synthetic voices, on Level 1 the ACCS synthesis performed worst for both voices, PL1 and PL2. Ivona™ and the HTS synthesis received the best results on the lowest level of accuracy/speed preset. The ACCS synthesis also had the worst recognition on the highest Level 7. The best performance was found for the HTS synthesis.

These high results of the HTS are probably achieved thanks to the fact that both, HTS and the ASR system, are based on the same technology. Therefore, speech output by the HTS provides a somehow better fit for the ASR system's decoder. Additionally, the HTS generates very natural prosody and renders virtually no mismatches at the segment boundaries.

Tab. 1.

Automatic recognition accuracy and speed (real time ratio) for 5 synthetic voices and a natural model voice, on 7 levels of accuracy/speed preset.

Accuracy [%]							
	1	2	3	4	5	6	7
Human	**84.1**	**86.6**	**90.3**	**90.3**	**90.3**	**90.3**	**90.6**
PL2	50.2	53.8	55.2	58.1	57.4	62.8	63.9
HTS	61.7	**71.8**	**74.4**	**79.8**	**79.4**	**81.9**	**82.7**
BOSS	56.3	57.0	61.7	67.1	68.6	71.8	72.6
PL1	**48.4**	**46.9**	**48.7**	**52.0**	**51.6**	**56.7**	**59.9**
Ivona	**65.7**	69.7	71.8	74.7	74.4	77.6	80.9
Real time factor [%]							
Human	**21.98**	**27.42**	**35.15**	**54.83**	**73.45**	**142.45**	**307.37**
PL2	35.44	43.60	62.24	104.00	135.02	262.83	567.55
HTS	28.99	35.05	**48.39**	77.94	101.36	203.19	495.76
BOSS	**27.91**	**34.84**	48.54	**75.42**	**99.74**	**200.65**	**435.31**
PL1	32.22	39.95	56.06	95.49	125.02	249.34	529.26
Ivona	**36.94**	**44.81**	**63.74**	**105.33**	**136.22**	**266.7**	**658.41**

Tab. 2:

Summary table of test results for 44 sentences (277 words) as reported by NIST; word correctness ("Corr"); error rates for word substitutions ("Sub"), deletions ("Del"), and insertions ("Ins"); total error rate ("Err"); and sentences with errors ("S.Err").

Level 1						
	Corr	**Sub**	**Del**	**Ins**	**Err**	**S.Err**
Human	**86.3**	**12.3**	**1.4**	**2.2**	**15.9**	**54.5**
ACCS (PL2)	**57.0**	**38.3**	4.7	6.9	49.8	88.6
HTS	65.3	26.4	**8.3**	**3.6**	38.3	**77.3**
BOSS	61.0	33.6	5.4	4.7	43.7	86.4
ACCS (PL1)	59.6	**38.3**	**2.2**	**11.2**	**51.6**	**93.2**
Ivona	**70.0**	**26.0**	4.0	4.3	**34.3**	86.4
Level 7						
	Corr	**Sub**	**Del**	**Ins**	**Err**	**S.Err**
Human	**91.7**	**7.6**	**0.7**	1.1	**9.4**	**40.9**
ACCS (PL2)	66.4	28.2	**5.4**	2.5	36.1	77.3
HTS	**83.4**	**13.7**	**2.9**	**0.7**	**17.3**	**56.8**
BOSS	73.3	22.0	4.7	**0.7**	27.4	72.7
ACCS (PL1)	**65.7**	**31.0**	3.2	**5.8**	**40.1**	**81.8**
Ivona	81.9	14.1	4.0	1.1	19.1	63.6

6.2 Semantic analysis

The text output of the ASR system was analysed for similarities and difference in word recognition on Level 7 of accuracy/speed preset.

It was found that in none of the speech materials (including natural speech) the following words were recognised: *zabiję* (Eng. *to kill*, 1st person singular, future tense), *Hajdukowie*, *Hajduków* (surname), *Majka* (first name). All these words were substituted by other words. Only the word "*zabiję*" was not included in the ASR dictionary. All the other words are present in the ASR dictionary and the system could recognise them, but it did not.

When it comes to the synthetic word analysis only, the system had problems recognising the following words and substituted them with different ones: *a* → na, k, h, adam; *broń* → broni, brak, brali, grunt; *nastawiane* → nastawione, nastawiony; *niej* → danieli, jakiej, mnie; *opryskliwe* → opryskliwie, opryskliwy; *się* → misie, niż, wisi; *państwa* → państwach; *pańswtwo* → państwa; *zajść* → zaś, znaleźć; *zakupię* → zakupie, kupił; *żądam* → żądamy, żądał, nierządem, zarządem.

The analysis of errors was also performed on the level of 44 chunks into which the police report was divided. On Level 7 the system correctly recognised the following chunks (sometimes with small errors): *przecinek* (Eng. *comma*), *oświadczam przecinek* (Eng. *I declare comma*), *mieszkam wraz z żoną oraz dziećmi* (Eng. *I live with my wife and children*), *nie słyszałem tego przecinek* (Eng. *I did not hear that comma*), *ostatni raz doszło do tego siódmego czerwca dwa tysiące ósmego roku* (Eng. *the last time it took place on the seventh of June two thousand and eight*).

The following sentences had the highest error rate recognition: *przebywa ich córka Majka Nowak* (Eng. *stays their daughter Majka Nowak*), *praktycznie codziennie u Państwa Hajduków* (Eng. *practically every day at the Hajdukowie*), *w którym zamieszkują Państwo Hajdukowie* (Eng. *in which the Hajdukowie live*), *moje córki przez Nowak* (Eng. *my daughters by Nowak*), *mówiąc jej przecinek że ją zabiję* (Eng. *telling her comma that I will kill her*), *i że zakupię broń w tym celu* (Eng. *and that I will buy a gun for that aim*).

7. Conclusions and discussion

In this paper, the evaluation of speech synthesis systems was performed using the automatic speech recognition system for Polish. The HTS system based on the Hidden Markov Models received the best results. The commercial unit selection Ivona™ text to speech system also scored very high, but the speed of recognition was the longest. The ACCS synthesis with MBROLA, a diphone concatenation system, had the worst results.

Human speech was recognised very well with the results slightly above 90% of accuracy for the Levels 3-7. However, the good recognition could result from the fact that Piotr's voice was already in the

Jurisdict database on which the acoustic models for the ASR system were trained. Although the speaker-adaptation was not applied for the recognition of Piotr's voice, the fact of having his voice in the Jurisdict database could have improved the recognition. For the future analysis, the recording of a speaker who is not in the Jurisdict database should be used for comparison.

The next step would be to perform the speaker-adaptation using the synthetic speech and see how the ASR performance changes, and if the improvement is close to the improvement when the human speech is input to the ASR system.

The high accuracy recognition of the synthetic speech for the HTS and the unit selection speech synthesisers show that these technologies can successfully be used in communicative situations, when human speech is replaced by synthetic voices and when the human ear is replaced by the automatic speech recognition system.

8. Acknowledgements

The authors are currently supported by grant "Collecting and processing of the verbal information in military systems for crime and terrorism prevention and control." (OR 00017012).

Bibliography

[1] Bachan, J. 2007. Automatic Close Copy Speech Synthesis. In: Speech and Language Technology. Volume 9/10. Ed. Grażyna Demenko. Poznań: Polish Phonetic Association. 2006/2007, pp. 107-121.

[2] Kuczmarski, T. 2010. HMM-based Speech Synthesis Applied to Polish. In: *Speech and Language Technology, vol 12/13*, Ed. Grażyna Demenko & Agnieszka Wagner, Poznań: Polish Phonetic Association, 2009/2010, pp. 221-228.

[3] Demenko, G., Klessa, K., Szymański, M. & Bachan, J. 2007. The design of Polish speech corpora for speech synthesis in BOSS system. In: *Proceedings of XII Sympozjum „Podstawowe Problemy Energoelektroniki, Elektromechaniki i Mechatroniki" (PPEEm'2007)*. Wisła, Poland, pp. 253-258.

[4] Osowski, Ł. & Kaszczuk, M. 2009. The IVO software Blizzard Challenge 2009 entry: Improving IVONA text-to-speech. In: *Blizzard Challenge Workshop*, Edinburgh, Scotland, September 2009.

[5] Ivona™ Text to Speech. <http://www.ivona.com>

[6] Demenko, G., Cecko, R., Szymański, M., Owsianny, M., Francuzik, P. & Lange, M. 2012. Polish speech dictation system as an application of voice interfaces. In: *Proceedings of 5th International Conference on Multimedia Communications, Services and Security. Communications in Computer and Information Science, Vol. 287*, Kraków 2012, pp. 68-76.

[7] Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vrecken O. 1996. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In: *Proceedings of ICSLP 96, vol. 3*. Philadelphia, pp.1393-1396.

[8] Szklanny, K. & Masarek, K. 2002. PL1 – A Polish female voice for the MBROLA synthesizer. Copying the MBROLA Bin and Databases.

[9] Bachan, J. 2011. *Communicative Alignment of Synthetic Speech*. Ph.D. Thesis. Institute of Linguistics. Adam Mickiewicz University in Poznań.

[10] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. & Tokuda, K. 2007. The HMM-based speech synthesis system (HTS) version 2.0. In: *Proceedings of 6th ISCA Workshop on Speech Synthesis (SSW-6)*. August 2007.

[11] Breuer, S., Stober, K., Wagner, P. & Abresch, J. 2000. Dokumentation zum Bonn Open Synthesis System BOSS II, Unveroffentliches Dokument, IKP, Bonn, <http://www.ikp.uni-bonn.de/>

[12] BOSS, the Bonn Open Synthesis System. <http://www.ikp.uni-bonn.de/forschung/phonetik/sprachsynthese/boss/>, accessed on 2010-09-19.

[13] Klessa, K. & Bachan, J. 2008. An investigation into the intra- and inter-labeller agreement in the JURISDIC database. In: *Speech and Language Technology. Volume 11*. Ed. Grażyna Demenko, Krzysztof Jassem & Stanisław Szpakowicz. Poznań: Polish Phonetic Association, pp. 47-54.

[14] Szymański, M. & Grocholewski, S. 2005. Transcription-based automatic segmentation of speech. In: *Proceedings of 2nd Language & Technology Conference*, Poznań, pp. 11–14.

Authors:

Jolanta Bachan, Ph.D.
Institute of Linguistics
Adam Mickiewicz University
al. Niepodległości 4
61-874 Poznań, Poland
tel. +48 502 544 279
fax +48 61-829-36-62
e-mail: *jolabachan@gmail.com*

Tomasz Kuczmarski, M.A.
Institute of Linguistics
Adam Mickiewicz University
al. Niepodległości 4
61-874 Poznań, Poland
tel. +048 509-079-785
e-mail: *tkucz@amu.edu.pl*

Piotr Francuzik, B.A.
Poznań Supercomputing and Networking Center
ul. Zwierzyniecka 20
60-814 Poznań, Poland
tel. +48 661-859-110
e-mail: *piotr.francuzik@speechlabs.pl*