Evaluation of Freely Available Speech Synthesis Voices for Halef

Martin Mory[1,2], Patrick Lange[1,3,4], Tarek Mehrez[1,5,6], David Suendermann-Oeft[1]

[1] [1]DHBW Stuttgart, Stuttgart, Germany
[2] [2]Hewlett-Packard, Böblingen, Germany
[3] [3]Linguwerk, Dresden, Germany
[4] [4]Staffordshire University, Stafford, UK
[5] [5]University of Stuttgart, Stuttgart, Germany
[6] [6]German University in Cairo, Cairo, Egypt
martin.mory@hp.com, patrick.lange@linguwerk.de, tarekmmehrez@gmail.com, david@suendermann.com

Abstract. We recently equipped the open-source spoken dialog system (SDS) Halef with the speech synthesizer Festival which supports both unit selection and HMM-based voices. Inspired by the most recent Blizzard Challenge, the largest international speech synthesis competition, we sought to find which of the freely available voices in Festival and those of the strongest competitor Mary are promising candidates for operational use in Halef. After conducting a subjective evaluation involving 36 participants, we found that Festival was clearly outperformed by Mary and that unit selection voices performed en par, if not better, than HMM-based ones.

1 Introduction

Spoken dialog systems (SDSs) developed in academic environments often substantially differ from those in industrial environments [1, 2] with respect to the data models (statistical vs. rule-based), the use cases (demonstration vs. productive usage), the underlying protocols, APIs, and file types (self-developed vs. standardized ones) and many more criteria.

To bridge the gap between both worlds, we have developed the SDS Halef (*H*elp *A*ssistant *L*anguage-*E*nabled and *F*ree) [3], being entirely based on open-source components, mainly written in Java. Unlike other academic SDSs, Halef's architecture is distributed, similarly to industrial SDSs; see Figure 1. In addition, Halef adheres to the following industrial standards:

- JSGF (Java Speech Grammar Format) as rule-based speech recognition grammar format [4],
- MRCP (Media Resource Control Protocol) [5] for exchange coordination of voice browser, speech recognition and synthesis components,
- RTP (Real-time Transport Protocol) [6] for audio streaming,

– SIP (Session Initiation Protocol) [7] for telephony, and
– VoiceXML [8] for dialog specification.

Simplifying the usual process for pursuing speech-driven applications, Halef provides a framework for implementing such applications by supplying the speech recognition and synthesis resources, leaving the dialog's logic as the only variable, to be controlled by the user.

The question answering system demonstrated in [3] is the first application being realized with Halef. The current version of Halef is limited to the English language, however, we are working on two additional applications which will be based on German (an information system for Stuttgart's public transportation authority and an intoxication checkup).

The quality of the speech synthesis component is crucial for the usability of a spoken dialog system. Commercially, this quality controls how appealing the system is for the ordinary user. Originally, Halef was based on the obsolete speech synthesizer FreeTTS [9] whose active development was terminated years ago and which caused user complaints about the synthesis quality. Our first step to improve quality was to enable our platform to use modern HMM-based voices. We did so by integrating the TTS system Festival [10] developed at the Centre for Speech Technology Research of the University of Edinburgh. This pushed Halef few steps forward towards a high-performing research-based SDS.

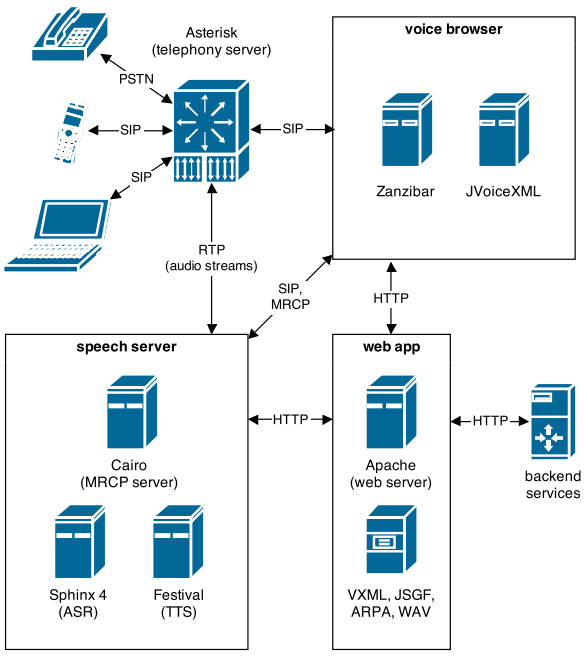

Fig. 1. High-level architecture of Halef

The decision to use Festival as the new TTS system was based on results of the last Blizzard Challenge [11], the world's most comprehensive speech synthesis evaluation. The results indicated that the HTS [12] voices of Festival were rated better than those of the other free TTS systems.

As it turned out, a number of HTS voices being accessible in the Festival online demo, and likely the ones achieving the best performance in the Blizzard challenge, are not publicly available. That is why the results and conclusions from the Blizzard Challenge cannot necessarily be applied to our setting. Therefore, we decided to conduct an own study to compare all those voices which are freely available to the public.

As indicated by the Blizzard challenge, another TTS system of prime quality is Mary [13], developed by the DFKI (German Research Center for Artificial Intelligence). We included Mary's English voices in our study as well. Table 1 provides an overview about the voices we compared in the evaluation. This set

voice	system	technique
cmu-bdl-hsmm	Mary	HSMM
cmu-rms-hsmm	Mary	HSMM
cmu-slt-hsmm	Mary	HSMM
cmu_us_awb_cg	Festival	Clustergen [14]
cmu_us_clb_arctic_clunits	Festival	cluster unit selection
cmu_us_rms_cg	Festival	Clustergen
cmu_us_slt_arctic_hts	Festival	HMM
dfki-obadiah-hsmm	Mary	HSMM
dfki-obadiah	Mary	unit selection
dfki-poppy-hsmm	Mary	HSMM
dfki-poppy	Mary	unit selection
dfki-prudence-hsmm	Mary	HSMM
dfki-prudence	Mary	unit selection
dfki-spike-hsmm	Mary	HSMM
dfki-spike	Mary	unit selection
kal_diphone	Festival	diphone unit selection

Table 1. Overview of the compared voices

of voices covers the different synthesis principles of unit selection as well as statistical parametric synthesis.

2 Experimental Setup

In order to get a reliable rating of the 16 freely available TTS voices, subjects had to assess samples of each of these voices, letting them rate their absolute quality on a six-level Likert scale. The levels from which to chose were 'utterly bad' (1), 'poor' (2), 'okay' (3), 'fine' (4), 'good' (5) and 'excellent' (6). Each voice had to be rated using exactly one of the named levels and independently of other voices.

The sentence "This is some example speech to evaluate the quality of available Festival and Mary voices" was used as the reference text being synthesized by all the 16 voices. The test audience consisted of persons who are familiar with speech processing and laypersons as well, ensuring that both specialist and non-specialist perceptions are taken into consideration.

As main measure to compare voices in this study, we calculated the mean opinion score (MOS) of the quality score described above. To test the significance of conclusions, we are using the Welch's t test since we are comparing independent samples expected to be Gaussian distributed. We assumed a significance level of 5%.

3 Results

Table 2 shows the results of the survey. For each voice and each level, the count of subjects who rated the voice with that score is given. The right column provides the MOS by which the table is sorted. The total number of participants was 36.

voice	(1)	(2)	(3)	(4)	(5)	(6)	MOS
dfki-spike	0	4	5	9	9	9	4.39
dfki-obadiah	1	4	6	10	11	4	4.06
dfki-spike-hsmm	0	5	7	11	9	4	4.00
cmu-bdl-hsmm	0	2	16	8	9	1	3.75
dfki-poppy	3	7	7	6	11	2	3.58
dfki-prudence	3	8	8	5	7	5	3.56
cmu-rms-hsmm	1	7	9	12	6	1	3.50
dfki-prudence-hsmm	1	7	13	8	5	2	3.42
cmu_us_slt_arctic_hts	1	6	15	8	6	0	3.33
cmu_us_rms_cg	1	7	16	6	5	1	3.28
cmu-slt-hsmm	2	11	9	9	3	2	3.17
dfki-poppy-hsmm	4	12	9	3	7	1	3.00
dfki-obadiah-hsmm	2	14	10	5	4	1	2.94
cmu_us_awb_cg	12	18	6	0	0	0	1.83
kal_diphone	20	13	1	0	2	0	1.64
cmu_us_clb_arctic_clunits	27	9	0	0	0	0	1.25

Table 2. Evaluation results, sorted by MOS

4 Discussion

4.1 Ranking

The voice dfki-spike achieved the highest MOS (4.39), but the Welch test was unable to confirm that it is significantly better than the voices dfki-obadiah and dfki-spike-hsmm. However, dfki-spike was found to be significantly better than

the voice cmu-bdl-hsmm and all the ones with worse MOS. Therefore, the first three voices can be interpreted as a cluster of the best rated voices in the study. The voices dfki-poppy, dfki-prudence, cmu-rms-hsmm, and dfki-prudence-hsmm are very close as well and form the second best-rated cluster. The third cluster is formed by the voices cmu_us_slt_arctic_hts, cmu_us_rms_cg, cmu-slt-hsmm, dfki-poppy-hsmm, and dfki-obadiah-hsmm, all of which do not show significant mean differences in the Welch test. The fourth cluster, consisting of the three worst voices of the study is concluded by the Arctic cluster unit selection voice.

4.2 Unit Selection vs. HSMM

The study included a number of voice pairs with the same speaker but different synthesis techniques. For each of the unit selection voices dfki-spike, dfki-obadiah, dfki-poppy, and dfki-prudence, there are HSMM-based counterparts. For dfki-poppy and dfki-obadiah, the unit selection version was rated significantly better than their HSMM-based counterparts. The two other pairs show the same tendence without being significantly different. Apparently, the natural sound of the unit selection voices was perceived to be more important for the test audience than the artifacts caused by the unit concatenation. This is an unexpected result, underlining that the evaluation of voice quality is very subjective and showing that unit selection can fulfill expectations in practical usage.

4.3 Festival vs. Mary

Our experiment shows that the results of the Blizzard Challenge [11] cannot be applied to our setting. Taking into consideration that the best rated Festival voice in our study is significantly worse than multiple Mary voices and furthermore that the three worst voices are Festival voices, it is unambiguous that Mary outperformed Festival based on the set of freely available voices. This is further evidenced by the MOS of Mary (3.58) versus that of Festival (2.27). The Festival unit selection voices clearly sound artificial, the accentuation is incorrect or missing at all and several sound distortions inhibit the understanding of the synthesized speech. Opposed to that, the Mary unit selection voices show only very few distortions and seem to have been recorded at higher sample rates which greatly improves the quality of the synthesis.

5 Conclusions

We presented results of a study comparing the synthesis quality of freely available voices of the speech synthesizers Festival and Mary. It turned out that Mary clearly outperforms Festival with the three best rated voices dfki-spike, dfki-obadiah, and dfki-spike-hsmm, because these voices are ahead of the Festival voices in the field in terms of better emphasis, less sound distortions and more natural sounding. Surprisingly, two of the three winners are unit-selection-based voices.

6 Future Work

Having identified the three best out of all voices available to us, we decided to equip Halef with the speech synthesizer Mary. To make sure that the present study's results are applicable to spoken dialog systems in operation, we will conduct a second evaluation round where subjects will rate full conversions with Halef rather than isolated recordings.

Acknowledgements

We thank all participants of our subjective evaluation.

Bibliography

[1] R. Pieraccini and J. Huerta, "Where Do We Go from Here? Research and Commercial Spoken Dialog Systems," in *Proc. of the SIGdial*, (Lisbon, Portugal), 2005.

[2] A. Schmitt, M. Scholz, W. Minker, J. Liscombe, and D. Suendermann, "Is It Possible to Predict Task Completion in Automated Troubleshooters?," in *Proc. of the Interspeech*, (Makuhari, Japan), 2010.

[3] T. Mehrez, A. Abdelkawy, Y. Heikal, P. Lange, H. Nabil, and D. Suendermann-Oeft, "Who Discovered the Electron Neutrino? A Telephony-Based Distributed Open-Source Standard-Compliant Spoken Dialog System for Question Answering," in *Proc. of the GSCL*, (Darmstadt, Germany), 2013.

[4] A. Hunt, "JSpeech Grammar Format. W3C Note," `http://www.w3.org/TR/2000/NOTE-jsgf-20000605`, 2000.

[5] D. Burnett and S. Shanmugham, "Media Resource Control Protocol Version 2 (MRCPv2)," `http://tools.ietf.org/html/rfc6787`, 2012.

[6] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobsen, "RTP: A Transport Protocol for Real-Time Applications," Tech. Rep. RFC 3550, IETF, 2003.

[7] A. Johnston, *SIP: Understanding the Session Initiation Protocol*. Norwood, USA: Artech House, 2004.

[8] M. Oshry, R. Auburn, P. Baggia, M. Bodell, D. Burke, D. Burnett, E. Candell, J. Carter, S. McGlashan, A. Lee, B. Porter, and K. Rehor, "VoiceXML 2.1. W3C Recommendation," `http://www.w3.org/TR/2007/REC-voicexml21-20070619`, 2004.

[9] W. Walker, P. Lamere, and P. Kwok, "FreeTTS: A Performance Case Study," tech. rep., Sun Microsystems, Santa Clara, USA, 2002.

[10] P. Taylor, A. Black, and R. Caley, "The Architecture of the Festival Speech Synthesis System," in *Proc. of the ESCA Workshop on Speech Synthesis*, (Jenolan Caves, Australia), 1998.

[11] M. Charfuelan, S. Pammi, and I. Steiner, "MARY TTS Unit Selection and HMM-Based Voices for the Blizzard Challenge 2013," in *Proc. of the Blizzard Challenge*, (Barcelona, Spain), 2013.

[12] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-Independent HMM-based Speech Synthesis System - HTS-2007 System for the Blizzard Challenge 2007," in *Proc. of the ICASSP*, (Nevada, USA), 2008.

[13] M. Schroeder and J. Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," *International Journal of Speech Technology*, vol. 6, no. 4, 2001.

[14] A. W. Black, "Clustergen: a statistical parametric synthesizer using trajectory modeling.," in *INTERSPEECH*, ISCA, 2006.