Evaluating Commercial and Open Source Text-to-Speech Synthesis Considering Specific Problem Classes

Felix Burkhardt

Telekom Innovation Laboratories
Felix.Burkhardt@telekom.de

Kurzfassung: Current state-of-the-art speech synthesizers for domain-independent systems still struggle with the challenge of generating understandable and natural-sounding speech. This is mainly because the pronunciation of words of foreign origin, inflections and compound words often can't be handled by rules and there are also too many of them for inclusion in exception dictionaries. We describe an approach to evaluate Text-to-Speech Synthesizers with a subjective listening experiment. The focus is to differentiate between known problem classes for speech synthesizers. We distinguish the following problem classes: Abbreviations, acronyms, acronym-abbreviations, addresses, compounds, dialectal expressions, exclamations, foreign origin words, German English (Denglisch), hetero-phonic homographs, named entities, inflected verbs, numbers, units and dates, rare words. We included some longer texts like short news feeds or e-mails. Because as a rule data-based speech synthesizers perform very different depending on how the target text fits to the data model, a large number of sentences must be tested in order to minimize the chance factor for the test sentences being part of the synthesizer's training data. Word lists for each of the above mentioned categories were compiled and synthesized by a commercial and an open source synthesizer, both being based on the non-uniform unit-selection approach. The synthesized speech was evaluated by a human judge using the Speechalyzer toolkit and the results are discussed. It shows that especially words of foreign origin were pronounced badly by both systems.

1 Introduction

In this study we are interested in two questions:

- In how far do known "problem classes" for speech synthesizers affect the quality of pronunciation?
- What's the difference with respect to the quality of pronunciation between a commercially developed synthesizer and an open source development?

Current state-of-the-art text to speech synthesizers for domain-independent systems that are based on the non-uniform unit-selection approach still struggle with the challenge of generating understandable and natural-sounding speech. This is mainly because the pronunciation of words of foreign origin, inflections and compound words often can't be handled by rules but there are also too many of them for exception dictionaries. Besides, even if the correct pronunciation would be known to the synthesizer, the necessary syllable combinations are often not present in the acoustic database, which leads to audible discontinues in the resulting output, especially with synthesizers based on non-uniform unit-selection.

Although the occurrence of each of these hard-to-pronounce words is very rare, the large number of the entirety of these words means that they occur in almost every sentence, the so-called "large number of rare events" phenomenon.

Many articles in the literature can be found on the evaluation of audio quality, not only focused on speech synthesis but even more general on speech transmission systems, codecs and others [6], [5].

A mean opinion scale (MOS) has been the recommended measure of synthesized speech quality [5]. Mostly the literature on speech synthesis evaluation is concerned with the best way to question the human listeners in subjective tests, [2], [4]. The ITU recommendation [5] suggests beneath "overall impression" the following categories; "listening effort", "comprehension problems", "articulation" and "acceptance". It also posits that at least five different sources of audio (or systems) should be used in these type of evaluations, including a "natural voice degraded with multiplicative noise".

While leaving the design of questionnaires out of the focus of this work and simply asking for a general mean opinion score (MOS), we focus on the text material that is the basis of the evaluation, like [1]. [8] used a short new article and e-mail as text material to cover their target domains. With respect to participants [2] differentiate three different groups, expert listeners, volunteers and paid participants.

The question which factors in a measurement are meaningful and independent and how to compute this has been thoroughly discussed in [9]. The authors write that studies should either follow the global approach, essentially asking for overall impression of sound quality, or the specific approach, i.e. comparing items representing different aspects of speech quality.

In this paper we use the first approach and neglect a more subtle analysis of the factors underlying one judgment. As speech tests are costly, we simply focus on the question which text material to include in the test.

The factors that influence the overall time to conduct a speech experiment are: the number of systems under comparison, the size of the text material, the number of listeners and the complexity of the questionnaire. Because manual speech evaluation is costly, not all these dimensions in a listening experiment can be large, The motivation of this study was to develop a fast and easy approach to evaluate Text-to-Speech synthesizers with a focus on some application scenario. In this paper the application is to find the problems that speech synthesizers might have with known problem categories.

The text material in this study is very large but the evaluation is based on a single annotator's judgments. If the approach will be used to compare speech synthesizers for a concrete application, definitely a larger number of labelers must be used, but then the text material can be of smaller size.

2 The Text material

Because by definition data-based speech synthesizers perform very different depending on how the target text fits to the data model, it follows that a large number of sentences should be tested in order to minimize the chance factor for the test sentences being part of the synthesizer's training data. Of course also the text material should stem from the domain of the target application in which the speech synthesizer will be used.

The text material in general evaluations should of course cover several domains, or what [2] call "genres". In this study we defined a set of "problem classes", i.e. short sentences or isolated words that included at least one case of a word that is known to cause problems for the pronunciation module in Text-to-Speech synthesis.

In Figure 1 the number of samples per problem class is displayed. As can be seen, it's by far not equally distributed. Some of the examples were simply thought up by the author, some, like the rare events data, were collected in text data collections.

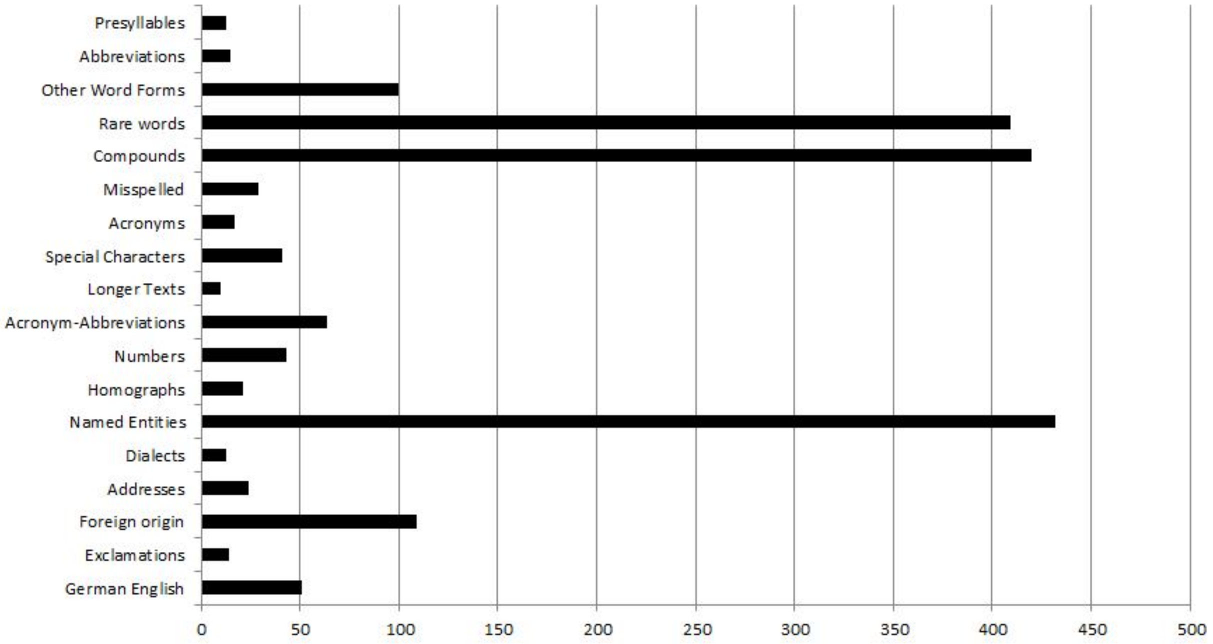


Abbildung 1 - Number of samples per problem class, for better comparison ordered like Figure 3

The following discusses each problem class in detail.

Abbreviations: Abbreviations should be expanded to the most common expressions, for example “etc.” to “et cetera”. The problem is, beneath that they have to be collected in a special dictionary, is that some of them have several meanings in different contexts.

Acronyms: Acronyms are a form of abbreviation but should be pronounced like one word and not spelled out as single letters, for example “NATO”.

Acronym-Abbreviations: These abbreviations should be pronounced as single letters, the difficulty is that some of them are, in the overwhelming contexts, English, one example would be “FBI”.

Addresses: Address formats use some conventions for abbreviations that are only valid in this context. Besides, converting addresses is a wide spread use case for Text-to-Speech synthesizers. We included street as well as internet or e-mail addresses.

Compounds: Compounds are words constructed by simple concatenation of other words and appear very often in German. The difficulty for Text-to-Speech synthesizers is, that at the (morpheme-) borders, pronunciation rules based on syllable sonority hierarchy fail because the word is still spoken like a series of words and a glottal stop should have been inserted at the border between the words, for example “Dekadenzerscheinung”. Also very often these words are stressed on the wrong syllable.

Dialects: In Internet blogs, forums and social networks, local user groups frequently use transcription of their local dialect. Of course these might be in the extreme just like a foreign language and no speech synthesizer can be expected to be able to pronounce the whole phoneme inventory of each local dialect variation. Nonetheless, the degree of intelligibility can be evaluated to estimate the degree of difficulties raised by these phenomena.

Exclamations and Onomatopoeia: When spoken speech is transcribed like in stories, blogs, or e-mails, often non-linguistic exclamations appear, because it's normal to frequently use them in everyday speech. If they get read by a speech synthesizer which is unprepared for proper pronunciation, the result may be quite confusing. A typical example might be "tss-tss".

Foreign origin: Words of foreign origin, naturally, don't follow German pronunciation rules which might lead to difficulties for the Letter to Sound rules. We also counted words of English origin if we felt that they are fully integrated into the German language and there is no adequate German translation, for example. "camper", "soft" or "software".

German English (Denglisch): A special situation comes from the growing number of English-stemming words, commonly known as the "Denglisch" phenomena. Stemming at least partly from English, they don't follow German pronunciation rules.

Heterophonic Homographs: "Heterophonic homographs" are difficult for speech synthesizers because they are spelled the same but pronounced differently, based on their meaning. In some cases this can be detected by a grammar based syntax parser, a "Part-of-Speech" parser, but not in any case. Their number is much smaller than for example. for English, but still they appear and mispronunciation causes confusion. A typical example would be "Spielende".

Longer texts: Beneath isolated words or short word groups, it makes also sense to test longer texts in order to evaluate a natural rhythm and give the chance to calculate pronunciation based on context information. The chosen texts might be typical for voice services, short e-mails or SMS reading.

Misspelled: The analysis of news items as well as personal messages showed that errors and misspellings occur frequently in texts. Although of course a perfect error correction (as done by humans based on context information) can't be expected from speech synthesizers, graceful recovery and handling of such situations in a way that the intention of the writer is still understandable would be desirable.

Named entities: Named Entities can be of origin from any language and therefore might be hard to pronounce for a Text-to-Speech synthesizer that uses German pronunciation rules. Nonetheless they are very important for applications like news reading where they appear frequently, not to mention that they may appear in personal messages. We used mainly an excerpt from international movie actor's names stemming from England, USA, France and India.

Numbers and units: Numbers and dates are very important to convey facts and can be a hard challenge as their pronunciation often depends on context, for example. in 12e-3, 1-4 and -2 the dash always has a different function. Introducing dates and measures adds complexity to this task.

Pre-syllables: In German, verbs and other words can be combined with a number of pre-syllables specifying the meaning. This might result in difficult pronunciation based on the correct syllable to be stressed. A typical example might be "weggegangen".

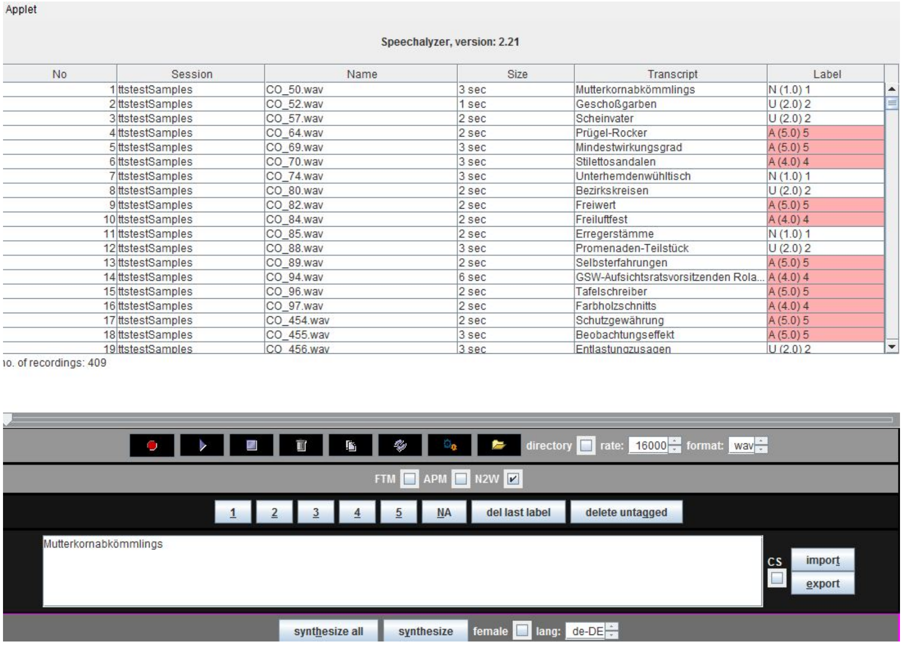
Special characters: Some special characters, for example the $ sign, are read, others, for example. brackets or hyphens, should be omitted.

Other words: A selection of words not necessarily fitting into the other categories, that are tricky to pronounce mainly because of uncommon phoneme combinations stemming from inflection. They mostly represent a collection of the author's experience when listening to news items read by speech synthesis.

Corpus Tokens: These words and word combinations were extracted from a news article corpus. They represent randomly selected items that occurred only once in the corpus, i.e. represent the "large number of rare events" phenomenon. They could indicate a realistic estimate of performance when reading newspaper articles.

3 Label process

Usually the studies that evaluate speech synthesis use more than just MOS tests, for example. include the typing of the utterances in order to test intelligibility of the system [2]. Due to cost and time restrictions, especially caused by the very high number of samples, we restricted this evaluation on only one five-point scale expressing "how natural sounds the pronunciation of this text sample?" and only one labeler (the author). The evaluation process was done in two big time frames, at first evaluating one synthesizer and some months later the other one in a series of sessions of about half an hour's length. The task was simplified by using the Speechalyzer toolkit [3] in combination with an Excel sheet. We realize that the value of a perception experiment based on only one recipient is not very high but wanted to get an estimate on the two main issues "problem classes" and "commercial vs. open-source system" as stated in the introduction.



Applet

Speechalyzer, version: 2.21

No	Session	Name	Size	Transcript	Label
1	ttstestSamples	CO_50.wav	3 sec	Mutterkornabkömmlings	N (1.0) 1
2	ttstestSamples	CO_52.wav	1 sec	Geschoßgarben	U (2.0) 2
3	ttstestSamples	CO_57.wav	2 sec	Scheinvater	U (2.0) 2
4	ttstestSamples	CO_64.wav	2 sec	Prügel-Rocker	A (5.0) 5
5	ttstestSamples	CO_69.wav	3 sec	Mindestwirkungsgrad	A (5.0) 5
6	ttstestSamples	CO_70.wav	3 sec	Stilettosandalen	A (4.0) 4
7	ttstestSamples	CO_74.wav	3 sec	Unterhemdenwühltisch	N (1.0) 1
8	ttstestSamples	CO_80.wav	2 sec	Bezirkskreisen	U (2.0) 2
9	ttstestSamples	CO_82.wav	2 sec	Freiwert	A (5.0) 5
10	ttstestSamples	CO_84.wav	2 sec	Freiluftfest	A (4.0) 4
11	ttstestSamples	CO_85.wav	2 sec	Erregerstämme	N (1.0) 1
12	ttstestSamples	CO_88.wav	3 sec	Promenaden-Teilstück	U (2.0) 2
13	ttstestSamples	CO_89.wav	2 sec	Selbsterfahrungen	A (5.0) 5
14	ttstestSamples	CO_94.wav	6 sec	GSW-Aufsichtsratsvorsitzenden Rola...	A (4.0) 4
15	ttstestSamples	CO_96.wav	2 sec	Tafelschreiber	A (5.0) 5
16	ttstestSamples	CO_97.wav	2 sec	Farbholzschnitts	A (4.0) 4
17	ttstestSamples	CO_454.wav	2 sec	Schutzgewährung	A (5.0) 5
18	ttstestSamples	CO_455.wav	3 sec	Beobachtungseffekt	A (5.0) 5
19	ttstestSamples	CO_456.wav	3 sec	Entlastungzusagen	U (2.0) 2

10. of recordings: 409

directory □ rate: 16000 format: wav

FTM □ APM □ N2W ☑

Mutterkornabkömmlings

CS □

synthesize all | synthesize | female □ lang: de-DE

Abbildung 2 - The Speechalyzer user interface used for the annotation task

The Speechalyzer was especially developed to ease the task to annotate or label large sets of audio data and was published as an open source project[1]. A screen shot of the interface is displayed in Figure 2. The synthesizers were interfaced by implementing special Interface classes for the framework. The Excel chart contained the text material and implements automatic im-

[1] https://github.com/dtag-dbu/speechalyzer

port and export (via the file system) to the Speechalyzer, as well as providing to generate the graphics and the computation of the mean result values.

4 Results

We tested our evaluation approach with two different Text-to-Speech systems; one by a commercial vendor and the open source Text-to-Speech system Mary developed by the DFKI [7]. For Mary, we used the latest stable version available in late 2014, namely version 5.0 with non-uniform unit-selection voice “dfki-pavoque-neutral”. We felt that this voice gives the best comparability to the commercial system, which also was based on non-uniform unit-selection.

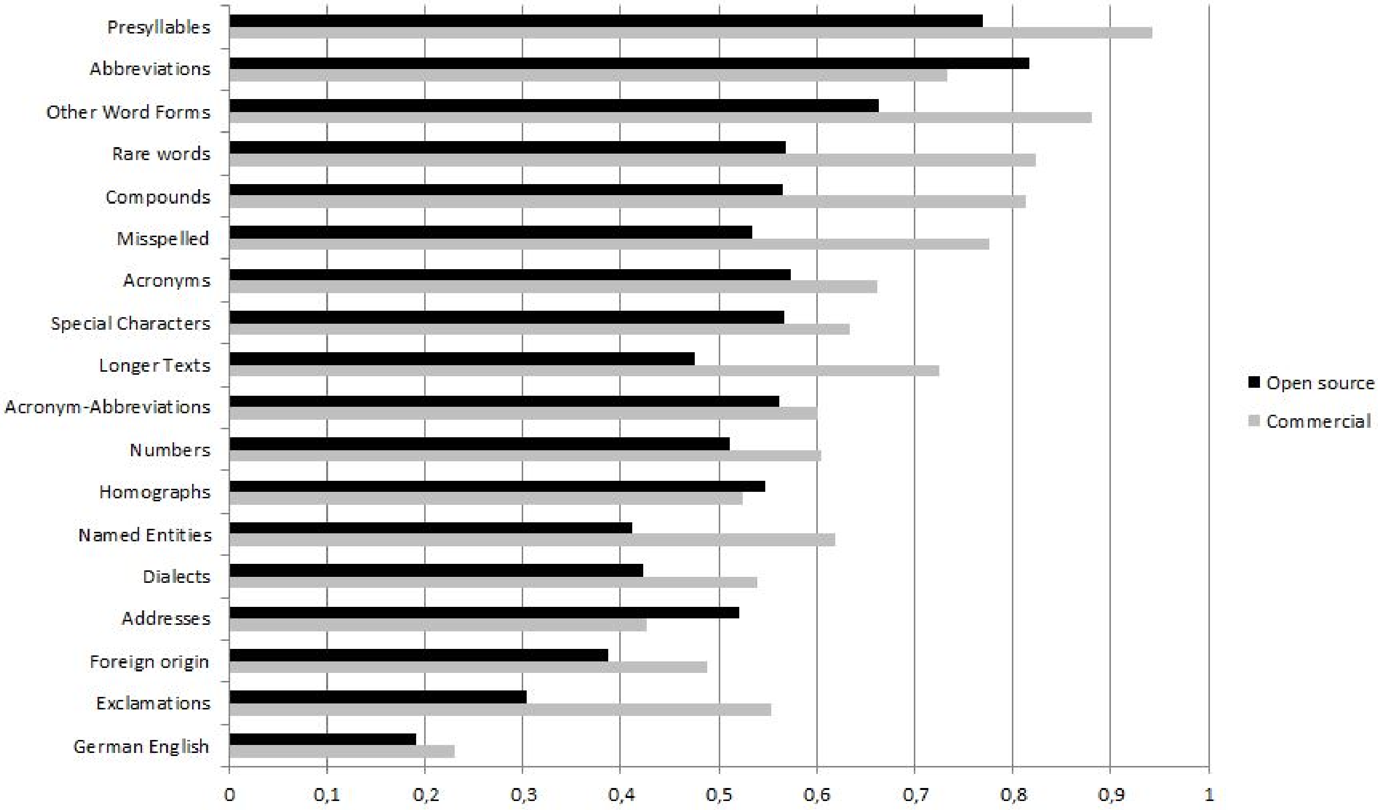


Abbildung 3 - Mean results for each evaluated system per problem class

With the commercially most successful approach to speech synthesis, nonuniform unit selection, best-fitting chunks of speech from large databases are concatenated, thereby minimizing a double cost function: best fit to neighbor unit and best fit to target prosody. Because signal manipulation is reduced as much as possible, the resulting speech sounds most natural (similar to the original speaker) as long as the utterance to synthesize is close to the original domain of the database.

Problems arise usually when unit combinations have to be synthesized that are under word level, i.e. shorter than single words. As the data is usually not recorded with a uniform pitch level, but the pitch movements are part of the diversity of the units, characteristic strange sounding pitch shifts appear in the output speech. This is most certainly one of the reasons that the problem class “German English” gives the worst performance for both systems, as the used words were most certainly not part of the original database.

The results of this evaluation are presented in figure 3. We projected the 1-5 Lickert scale on a 0-1 dimension, the values denote the arithmetic mean values of the sample judgments.

4.1 Comparison of commercial vs. open-source synthesizer

As expected, the commercial system outperforms the open source system in nearly all problem classes. In only two classes, names abbreviations, addresses and homographs, the open source systems delivers clearly better performance than the commercial version. We feel that this is mainly caused by a chance effect as the sample number of this two problem classes may not be large enough.

Companies can spend more money on labour to compile exception dictionaries and larger sample databases, so especially the much better performance for compounds, named entities and rare words is not a surprise.

4.2 Performance with respect to problem classes

The systems show a high correlation with respect to the problem classes. The words that have pre-syllables (mostly verb forms, for example "niedergeredet") show to cause the least problems, followed by abbreviations, rare words and compounds. But already compounds have only an 80 % success rate which means the two out of ten are badly pronounced. Dialectal expressions and exclamations are very unpredictable and unclear, so a bad value for these classes is not a surprise.

But the very low values for German-English and Foreign-Origin words show that the task to pronounce words that are not native German has to be tackled by Text-to-Speech synthesizers as they appear frequently and with rising probability in modern German.

5 Conclusions and Outlook

We presented an approach to evaluate Text-to-Speech systems manually based on known problem categories. The approach was used on two distinct systems, one being a commercial synthesizer and the other the open source synthesis system Mary. Overall the commercial synthesizer showed clearly a better performance which was to be expected given that quite a large team works on the synthesizer performance while the open source system usually gives a starting point but is meant to be improved by the users.

All in all the high number of pronunciation errors for both systems shows that there is still a long way to go to achieve results with a Text-to-speech synthesizer reading unrestricted content that can compare to a trained human speaker. Practical experience by the author (when getting read news RSS feeds on the way to work for some weeks) showed that one error per sentence is sufficient to impede a positive listening experience.

There are many flaws in this experiment, probably the most grace one is that the evaluation is based on a single annotator's judgments. If the approach will be used to compare speech synthesizers for a concrete application, definitely a larger number of labelers will be used.

Literatur

[1] BENOIT, C., M. GRICE und V. HAZAN: *The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences*. Speech Communication, Volume 18(4):381–392, June 1996.

[2] BLACK, A. W. und K. TOKUDA: *The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets*. In: *in Proceedings of Interspeech 2005*, S. 77–80, 2005.

[3] BURKHARDT, F.: *Fast Labeling and Transcription with the Speechalyzer Toolkit*. Proc. LREC (Language Resources Evaluation Conference), Istanbul, 2012.

[4] HINTERLEITNER, F., C. NORRENBROCK, and S. MÖLLER: *Perceptual quality dimensions of text-to-speech systems in audiobook reading tasks*. In *Elektronische Sprachsignalverarbeitung (ESSV 2013)*, pp. 44–49, mar 2013.

[5] ITU-P85: *Telephone transmission quality subjective opinion tests. a method for subjective performance assessment of the quality of speech voice output devices.*, 1994.

[6] RIX, A. W., J. G. BEERENDS, D.-S. KIM, P. KROON, and O. GHITZA: *Objective assessment of speech and audio quality&# 8212; technology and applications*. Audio, Speech, and Language Processing, IEEE Transactions on, 14(6):1890–1901, 2006.

[7] SCHRÖDER, M. and J. TROUVAIN: *The german text-to-speech synthesis system mary: A tool for research, development and teaching*. International Journal of Speech Technology, 6:365–377, 2003.

[8] SONNTAG, G. P., T. PORTELE, F. HAAS, and J. KÖHLER: *Comparative evaluation of six german tts systems*. In *In Proceedings of the European Conference on Speech Communication and Technology*, pp. 251–254, 1999.

[9] VISWANATHAN, M. and M. VISWANATHAN: *Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale*. Computer Speech & Language, pp. 55–83, 2005.