A Taxonomy of Specific Problem Classes in Text-to-Speech Synthesis: Comparing Commercial and Open Source Performance

Felix Burkhardt and Uwe D. Reichel
Telekom Innovation Laboratories, Research Institute for Linguistics, Hungarian Academy of Sciences
Winterfeldstr. 21, 10785 Berlin, Germany, 1068 Budapest VI., Benczur u. 33., Hungary
Felix.Burkhardt@telekom.de, uwe.reichel@nytud.mta.hu

Abstract

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 cannot be handled by rules. Furthermore there are too many of these for inclusion in exception dictionaries. We describe an approach to evaluating text-to-speech synthesizers with a subjective listening experiment. The focus is to differentiate between known problem classes for speech synthesizers. The target language is German but we believe that many of the described phenomena are not language specific. We distinguish the following problem categories: Normalization, Foreign linguistics, Natural writing, Language specific and General. Each of them is divided into five to three problem classes. Word lists for each of the above mentioned categories were compiled and synthesized by both a commercial and an open source synthesizer, both being based on the non-uniform unit-selection approach. The synthesized speech was evaluated by human judges using the Speechalyzer toolkit and the results are discussed. It shows that, as expected, the commercial synthesizer performs much better than the open-source one, and especially words of foreign origin were pronounced badly by both systems.

Keywords: text-to-speech, evaluation, problem-classes

1. Introduction

In this study we are interested in two questions:

- To what degree do known "problem classes" for speech synthesizers affect the quality of pronunciation?
- What is 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.

Nonuniform unit selection is the commercially most successful approach to speech synthesis. It works basically by concatenating best-fitting chunks of speech from large databases, 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.

Many errors occur because the pronunciation of inflections, compound words and words of foreign origin as well as so-called "non-standard words" (Sproat et al., 2001) often cannot be handled by rules. However, there are too many of these for inclusion in exception dictionaries. Furthermore, even if the correct pronunciation would be known to the synthesizer, the necessary syllable combinations are often not present in the acoustic database, leading to audible discontinuities 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 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 (Rix et al., 2006), (ITU-P85, 1994).

A mean opinion scale (MOS) has been the recommended measure of synthesized speech quality (ITU-P85, 1994). Mostly the literature on speech synthesis evaluation is concerned with the best way to question the human listeners in subjective tests, (Black and Tokuda, 2005), (Hinterleitner et al., 2013). The ITU recommendation (ITU-P85, 1994) suggests in addition to "overall impression" the following categories; "listening effort," "comprehension problems," "articulation" and "acceptance." It also posits that at least five different sources of audio 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, such as (Benoit et al., 1996). (Sonntag et al., 1999) used a short new article and e-mail as text material to cover their target domains. With respect to participants (Black and Tokuda, 2005) differentiate three different groups, expert listeners, volunteers and paid participants.

The question as to which factors in a measurement are meaningful and independent and how to compute this has been thoroughly discussed in (Viswanathan and Viswanathan, 2005). The authors write that studies should either follow the global approach, essentially asking for overall impression of sound quality, or the specific ap-

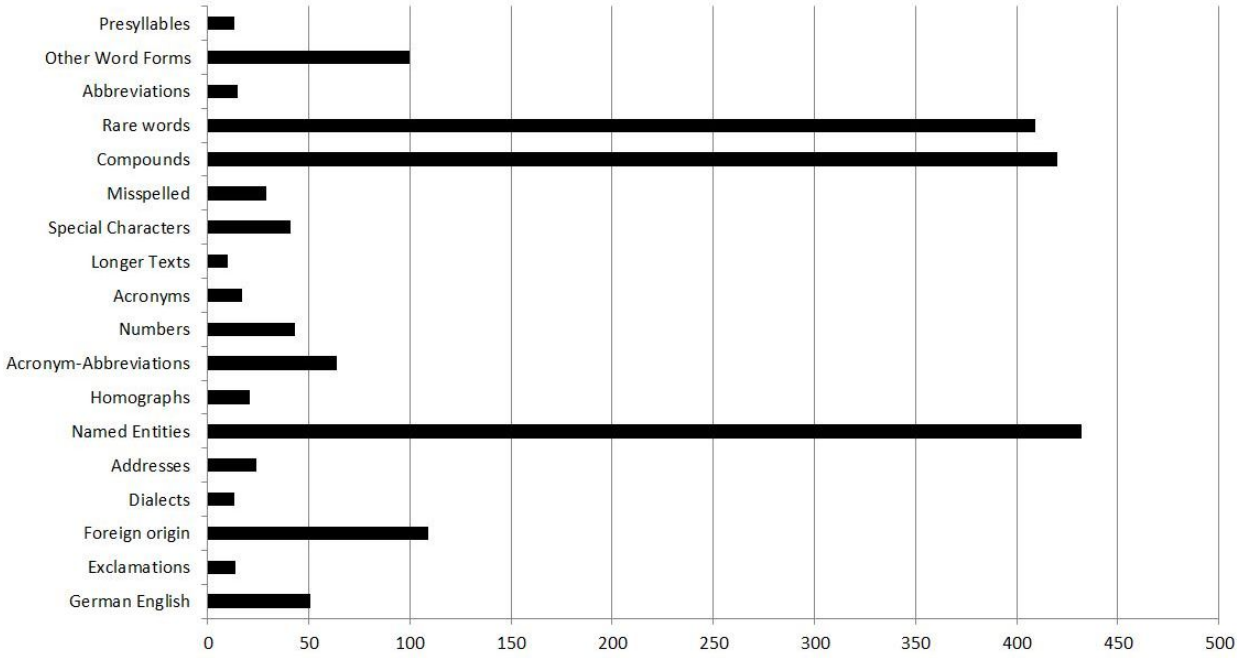

Figure 1: Number of samples per problem class, for better comparison ordered like Figure 2

proach, i.e. comparing items representing different aspects of speech quality.

The target language is German but we believe that many of the described phenomena are not language specific.

The text material in this study is quite extensive but the evaluation is based on the judgments of only two resp. one annotator. As (Wester et al., 2015) show, it's important to use a high number of human judges for subjective evaluation of text-to-speech synthesis. Due to the large size of the text material it was not possible in this study to fulfill this requirement, so the results can only be interpreted with caution. If the approach described here is to be implemented to compare speech synthesizers for a concrete application, definitely a larger number of labelers must be used. In this case the text material can be of smaller size.

One way to solve the problem of costly subjective listener tests for large test data sets is described in (Chevelu et al., 2015) where the authors propose to filter the acoustically most different audio files.

(Sproat et al., 2001) did something quite similar to the work done in this paper by developing quite an extensive taxonomy for non-standard words that are problematic for the letter-to-sound module. It consists of three groups; four categories of alphabetics (ALPHA), 13 categories of numerical phenomena (NUMBERS) and and six further categories (MISC).

Motivated by our experience with integrating text-to-speech synthesis into HCI (human computer interaction) systems, we suggest a rather different taxonomy which is described in the next section.

2. A Taxonomy of Problem Classes

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 (Black and Tokuda, 2005) identify as "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, the distribution is far from being equal. Some of the examples were created at random by the author, some, like the rare events data, were collected in text data collections.

The following discusses each problem class in detail.

2.1. Normalization

This group bundles all problem classes that deal with symbol resolution.

Abbreviations: Abbreviations should be expanded to the most common expressions, for example "etc." to "et cetera". The problem is, in addition to their having to be collected in a special dictionary, some of these 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". We did not look at abbreviations of normal words, e.g. "fplc" for "fireplace" like (Sproat et al., 2001) as we felt that this is not very common in German.

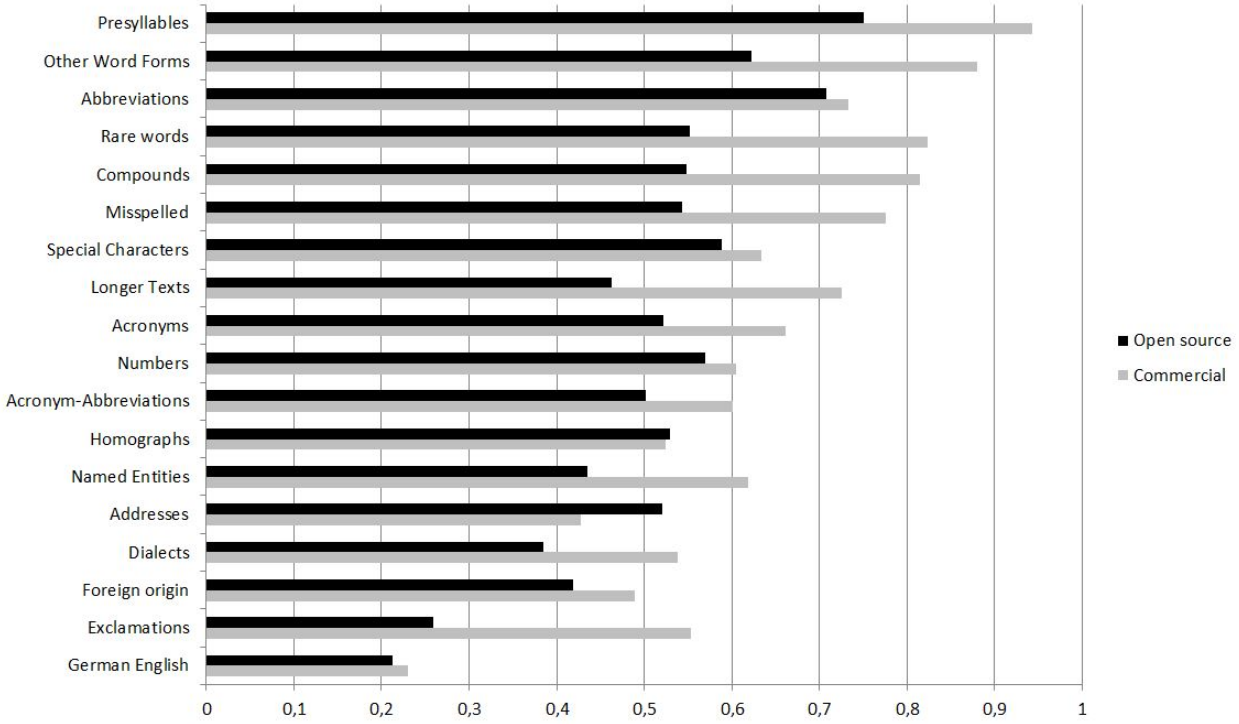

Figure 2: Mean results for each evaluated system per problem class

Acronym-Abbreviations: These abbreviations should be pronounced as single letters, the difficulty is that some of these 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. Furthermore, converting addresses is a wide spread use case for text-to-speech synthesizers. We included street as well as Internet or e-mail addresses.

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.

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

2.2. Foreign linguistics

This group deals with problems coming from words containing phonemes that don't exist in the high level language, in this case "standard German".

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 naturalness can be evaluated to estimate the degree of difficulties raised by these phenomena.

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".

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.

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

2.3. Natural writing

A group for phenomena resulting from the way people use text language naturally.

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.

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.

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

2.4. Language (German) specific

This group encompasses the problems that are specific to a certain language because they are typical for the linguistic structure of this language, in this case German.

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". Very often these words get stressed on the wrong syllable.

Prefixes: In German, verbs and other words can be combined with a number of prefixes specifying the meaning. This might result in difficult pronunciation based on the correct syllable to be stressed. A typical example might be "weggegangen". Some prefixes, e.g. "weg" must be stressed, while "ge" can't be stressed.

2.5. General

A final group for other classes.

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, in the English language, but still they appear and mispronunciation causes confusion. A typical example would be "Spielende."

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. One example would be "Scheinvater" ("mock-father").

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. One example would be "fahrradähnlichen" ("like a bicycle").

3. Label process

Usually the studies that evaluate speech synthesis use more than just MOS tests, but, for example, include the typing of the utterances in order to test intelligibility of the system (Black and Tokuda, 2005). 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 is the pronunciation of this sample?", with 1 for "*very badly pronounced*" and 5 for "*very natural pronunciation*".

The samples for the commercial synthesizer were labeled by only one labeler. As we were interested to get some insight into inter-labeler agreement, the samples for the open-source synthesizer were also labeled by a second labeler. Both labelers were expert listeners, i.e. trained phoneticians.

Overall the agreement is quite good (p value 0.000 for correlation coefficients after Spearman (0.64). Cohen's kappa value 0.78), only for the classes Addresses (p 0.067, Spearman coefficient 0.31, kappa 0.71) and Misspelled (p 0.101, Spearman coefficient 0.24, kappa 0.71) there is only a tendency for agreement.

To account for the ordinal level of measurement of the 5-level scale we calculated weighted kappa values, i.e. judgment mismatches were weighted by their absolute difference.

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 (Burkhardt, 2012) in combination with an Excel sheet.

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 3. The synthesizers were interfaced by implementing special Interface classes for the framework. The Excel chart contained the text material and implements automatic import 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

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

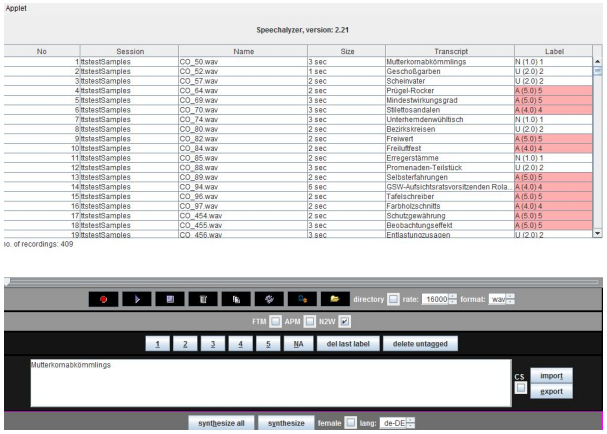

Figure 3: The Speechalyzer user interface used for the annotation task

open source text-to-speech system Mary developed by the DFKI (Schröder and Trouvain, 2003). 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.

Nonuniform unit selection is the commercially most successful approach to speech synthesis. It works basically by concatenating best-fitting chunks of speech from large databases, 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 2. We projected the 1-5 Likert scale (1 for "*very badly pronounced*" and 5 for "*very natural pronunciation*") 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.

We compared the performance of the two systems separately for each category by two sided Wilcoxon signed rank tests for paired samples with a significance level $\alpha = 0.05$. Type I error was corrected by controlling the False Discovery Rate as proposed by (Benjamini and Hochberg, 1995): the k-th lowest of n p-values must be below $\frac{k}{n} \cdot \alpha$ in order to indicate a significant difference. Overall, the commercial system outperforms the open source system which is expressed in significantly higher judgments in the seven categories "*Compounds, Corpus Tokens, Exclamations/Onomatopoeia, Longer texts, Misspelled, Named entities, and Other word forms*" [*]. The open source system however shows slight but not significant advantages for the three categories " *Addresses, Foreign origin, and Numbers and units*" [**].

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 (Spearman rho = 0.77, p=0.0002). 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 even the mean number of compounds that were rated less than .5 is 13% for the commercial system and 39% for the open-source system, which means that at least one word out of ten is badly pronounced. This might already prevent an acceptable user experience for applications that feature speech synthesis for unlimited domains.

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 a taxonomy of problem classes for text-to-speech synthesizers and used this in a text-to-speech system evaluation. 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 of the first 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.

6. Acknowledgements

The work of the second author is financed by a Feodor Lynen grant of the Alexander von Humboldt society.

7. Bibliographical References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, (1):289–300.

Benoit, C., Grice, M., and Hazan, V. (1996). 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.

Black, A. W. and Tokuda, K. (2005). The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common datasets. In *in Proceedings of Interspeech 2005*, pages 77–80.

Burkhardt, F. (2012). Fast labeling and transcription with the speechalyzer toolkit. *Proc. LREC (Language Resources Evaluation Conference), Istanbul*.

Chevelu, J., Lolive, D., Maguer, S. L., and Guennec, D. (2015). How to compare tts systems: A new subjective evaluation methodology focused on differences. *Proceedings Interspeech 2015*.

Hinterleitner, F., Norrenbrock, C., and Möller, S. (2013). Perceptual quality dimensions of text-to-speech systems in audiobook reading tasks. In *Elektronische Sprachsignalverarbeitung (ESSV 2013)*, pages 44–49, mar.

ITU-P85. (1994). Telephone transmission quality subjective opinion tests. a method for subjective performance assessment of the quality of speech voice output devices.

Rix, A. W., Beerends, J. G., Kim, D.-S., Kroon, P., and Ghitza, O. (2006). Objective assessment of speech and audio quality&# 8212; technology and applications. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):1890–1901.

Schröder, M. and Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377.

Sonntag, G. P., Portele, T., Haas, F., and K"ohler, J. (1999). Comparative evaluation of six german tts systems. In *In Proceedings of the European Conference on Speech Communication and Technology*, pages 251–254.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., and Richards, C. (2001). Normalization of non-standard words. *Computer Speech & Language*, 15:287–333.

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

Wester, M., Valentini-Botinhao, C., and Henter, G. E. (2015). Are we using enough listeners? no! an empirically-supported critique of interspeech 2014 tts evaluations. In *Proceedings Interspeech 2015*.