HMM-based Speech Synthesis from Audio Book Data
B002168

Kathrin Haag

Master of Science
Speech and Language Processing
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
2011

Abstract

In contrast to hand-crafted speech databases, which contain short out-of-context sentences in fairly unemphatic speech style, audio books contain rich prosody including intonation contours, pitch accents and phrasing patterns, which is a good pre-requisite for building a natural sounding synthetic voice. The following paper will give an overview of the steps that are involved in building a synthetic voice from audio book data.

After an introduction to the theory of HMM-based speech synthesis, the properties of the speech database will be described in detail. It will be argued that it is necessary to model specific properties of the database, such as higher pitched speech or questions, to achieve a better quality synthetic voice. Furthermore, the acoustic modelling of these properties will be explained in detail. Finally, the synthetic voice is evaluated on the basis of an online listening test.

Table of Contents

1 Introduction and motivation

Statistical parametric speech synthesis based on HMMs is one of the most studied speech synthesis approaches (cf. Tokuda et al. 2002, Yamagishi et al. 2008, Zen et al. 2009) and can generate fairly natural sounding speech. In contrast to unit-selection speech synthesis systems, HMM-based speech synthesis is very flexible with regard to speech modelling. While synthesised speech from unit-selection systems cannot be modified, and the output sounds basically like the speech from the input database, speech parameters can be manipulated in HMM-based speech synthesis. This is particularly interesting for speech appliances which require different kinds of speaking styles, such as audio books.

State-of-the-art speech synthesis systems produce natural sounding and intelligible speech, which is especially helpful for people with visual handicaps. Text-to-speech systems that read aloud texts from web sites or text documents are practicable for the facilitation of day-to-day tasks. However, when it comes to applets that are used for leisure, for example e-book readers, people do not only want a voice they understand well, but one which is pleasant to listen to. State-of-the-art e-book readers like the one that is built into Amazon's Kindle do not satisfy this condition yet. As a self-conducted internet research revealed, a large number of users regard Kindle's voice as natural, but also as a nuisance when they listen to it for a longer time because it speaks with a fairly monotone voice.

Using audio books as a speech database is a good means to get around this problem. In contrast to hand-crafted speech databases, which contain short out-of-context sentences in quite unemphatic speech style, audio books contain rich prosody including intonation contours, pitch accents and phrasing patterns, which is a good pre-requisite for building a natural sounding synthetic voice. Furthermore, audio books contain various speaking styles, which can be modelled and built into the synthetic voice so that an e-book reader can employ these styles while reading aloud a book text. How this can be accomplished will be presented in this paper. The outline of the paper will be as follows.

Chapter 2 will give an overview of the theory behind HMM-based speech synthesis. HMM-based speech synthesis will be explained in general, and on the basis of a training script for the HTS speech synthesis system that was developed at the University of Edinburgh. Chapter 3 will describe the nature of the audio book data in terms of a phonetic and prosodic analysis, and highlight the advantages and disadvantages of using audio book data for speech synthesis. Chapter 4 will outline the changes that were made to improve the speech synthesis system. Finally, chapter 5 will present a statistical evaluation of the synthetic voice, which is

based on the results of an online listening test.

2 The theory behind HMM-based speech synthesis

The theory behind HMM-based speech synthesis[1] will be explained on the basis of the HTS speech synthesis system, which was publicly released in 2002 and is undergoing a constant development (cf. http://hts.sp.nitech.ac.jp*)*. HTS is an extension of the Hidden Markov Model Toolkit (HTK) for automatic speech recognition (cf. http://htk.eng.cam.ac.uk*)* and uses hidden Markov models to model speech parameters. The training script that was used in this dissertation project was developed by Junichi Yamagishi, member of the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, in 2010. Training was conducted on the Edinburgh Compute and Data Facility (ECDF) (cf. http://www.ecdf.ed.ac.uk), which is a high performance cluster of servers and storage. The advantage of using this cluster is that software can be run in parallel on separate CPUs, which speeds up processes like feature extraction or tree-based context clustering considerably. This way, several large voices can be built and stored on the cluster, which is not possible on a standalone machine with limited amount of memory and storage space. In the following, the background of HMM-based speech synthesis and the steps that are followed in building the voice will be explained in more detail.

HMM-based speech synthesis is a statistical parametric speech synthesis approach. Compared to unit selection speech synthesis, which concatenates pre-recorded chunks of speech with minimal application of signal processing, HMM-based synthesis can be understood as generating the average of similar sounding speech units in the database (cf. Zen et al. 2009: 1040). In the framework, spectral, excitation and duration parameters are statistically and simultaneously modelled using HMMs. These parameters corresponding to input text are directly generated from HMMs themselves. The resulting synthetic speech may sound buzzy, but it is very smooth and stable compared to a unit selection voice. In addition to that, the speech units can be modified, which is not the case in concatenative synthesis where the generated speech is limited to the speech style in the recorded database. This is particularly problematic when the recorded speaker has a non-consistent speech style and shows large variations in e.g. prosody or the expression of emotions. When the database is very small, the synthetic output from unit selection will sound quite unnatural. A substantial

1 In fact, HSMMs (hidden semi-Markov models) are usually used in speech synthesis. These imply an explicit duration model, which allows for non-geometrical and non-exponential distributions of durations (cf. Yu and Kobayashi 2001: 235). The explicit duration model is not Markov, but the state transitions are, which is why it is called semi-Markov (cf. King 2010: 7).

advantage of HMM-based synthesis is that speech synthesis is much more flexible. Since all the speech parameters are statistically modelled within the framework of HMMs, many model adaptation and model interpolation methods can be adopted to control the model parameters and diversify the characteristics of generated speech (cf. Yamagishi et al. 2009: 1208). Moreover, very good synthesis results can be obtained with a comparatively small speech database.

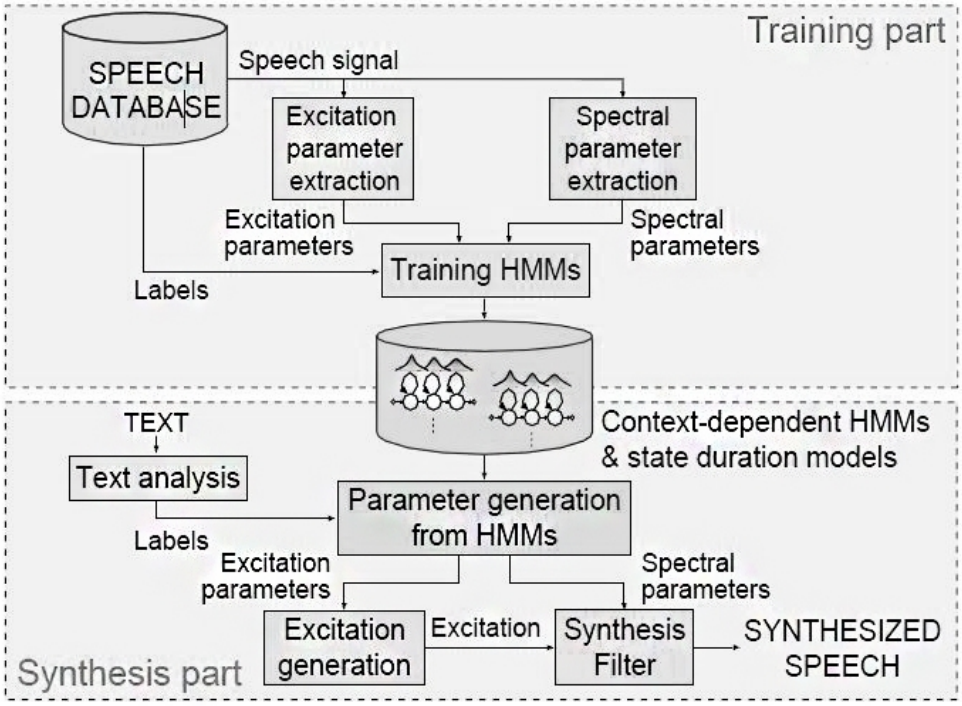

Fig. 1: Processes involved in HMM-based speech synthesis (cf. Tokuda 2009)

2.1 Feature extraction

The core architecture of a typical HMM-based synthesis system consists of a training part and a synthesis part, as shown in Figure 1. In the training step, excitation parameters (log f0 and band-limited aperiodic features for mixed excitation) and spectral parameters (39 Mel-cepstral coefficients) are extracted from a speech database. For f0 extraction, a voting of instantaneous frequency amplitude spectrum (IFAS) (cf. Arifianto et al.: 2004), fixed-point analysis (TEMPO) (Kawahara et al.: 1999) and the ESPS get_f0 tool (cf. Talkin: 1995) is used to distinguish between voiced and unvoiced signals. IFAS is a method that quantitatively determines the difference between voiced and unvoiced speech signals by measuring the degree of periodicity of a speech signal. This method has been proven to be extremely efficient in noisy environments and outperformed TEMPO and the ESPS get_f0 tool in a study by Arifianto and Kobayashi (2005). It is therefore useful to apply IFAS in f0 extraction of the audio book data, which was not recorded in a sound studio and is not completely noise-free. However, a voting method that combines all of these techniques was found to "reduce

errors such as f0 halving and doubling, and voiced/unvoiced errors" (Yamagishi et al.: 2008), and therefore a combination of these methods can be better than a single pitch tracker.

2.2 Embedded training

The extracted parameters are first modelled by context-independent monophone HMMs, and then embedded training is conducted. Embedded training is an iterative process, which is performed by the Expectation Maximization (EM) algorithm to align wave files and their corresponding transcriptions (cf. Jurafsky and Martin 2009: 221). No initial segmentations or phone durations are needed. The input to the algorithm are a wave file and its phonetic transcription, modelled by an HMM. The EM algorithm sums over all possible segmentations of words and phones and aligns the phonetic labels with the cepstral features extracted from the waveform. The model parameters λ of the HMMs are set so as to maximise the probability of the training data O given a set of word sequences W that corresponds to O (cf. Zen et al. 2009: 1042):

$$\lambda = \arg\max_{\lambda} \{p(O \mid W, \lambda)\}$$

The EM algorithm is initialised with a *flat start*. In a flat start initialisation, probabilities for back transitions to earlier phones are set to zero, while all other transitions are equally probable. The mean and variance are identical for each Gaussian, which makes computation very simple (cf. Jurafsky and Martin 2009: 360). The algorithm consists of two steps, namely the expectation step, in which the state occupation probabilities of being in state j at time t are estimated, and the maximization step, which utilizes these state occupation probabilities to re-estimate the HMM parameters. In the HTS training script, the algorithm iterates 5 times and the alignment of phonetic labels and cepstral features is re-estimated in each iteration. This way, unlikely alignments are replaced with more probable alignments and a more consistent and accurate set of labels is attained that we can train our models on. After embedded training is applied for monophone HMMs, these are converted to full-context HMMs, and embedded training is applied again. Full-context HMMs are context-dependent models, which in addition to phonetic quinphone contexts also model the segmental, prosodic and linguistic context of a speech unit (cf. Zen and Gales 2011: 4560).

2.3 Context-clustering

However, not all possible contexts can be captured by the context-dependent HMMs. If we consider a system with 40 phones plus silence, then there are 41^5 = 115,856,201 logical quinphone combinations. In a real corpus of English, not all of these combinations do actually occur, but it can be assumed that there are still a large number of quinphone combinations, which would require a high amount of training data to enable a robust estimation of the model parameters and to ensure that all possible quinphone combinations are covered (cf. Renals 2011: 15). Training on a sufficiently large speech database is computationally expensive, but we can compensate for unseen quinphones by applying context clustering, which enables us to model quinphones that were not observed in the training data. In addition to the phonetic quinphone contexts, the segmental, prosodic and linguistic contexts have to be considered as well. Since it is impossible to cover all possible contexts with a limited amount of training data, context clustering is applied for mel-cepstral coefficients, log f0, band-limited aperiodic features and state durations. With context clustering, every unseen context can be modelled.

Context clustering is a data-driven, top-down approach, which uses binary decision trees to cluster similar context-dependent HMM states into the same context classes (cf. Jurafsky and Martin 2009: 381). At the root of the tree, all states are shared. Yes/no questions split the pool of states, and the resultant state clusters are given by the leaves of the tree. Example questions for a phonetic decision tree include, “Is the left context a nasal?” or ‘”Is the right context a central stop?”. The questions at each node are chosen from a large set of predefined questions, and at each node the question that maximizes the likelihood of the data given the state clusters is chosen. The likelihood of a state cluster, assuming Gaussian distributions is given as (cf. Renals 2011):

$$L(S) = \sum_{i=1}^{K} \log P(X_i \mid \mu_S, \sum_S)$$

Splitting is stopped if either the improvement given by the splitting question or the amount of data associated with a split node falls below a given threshold.

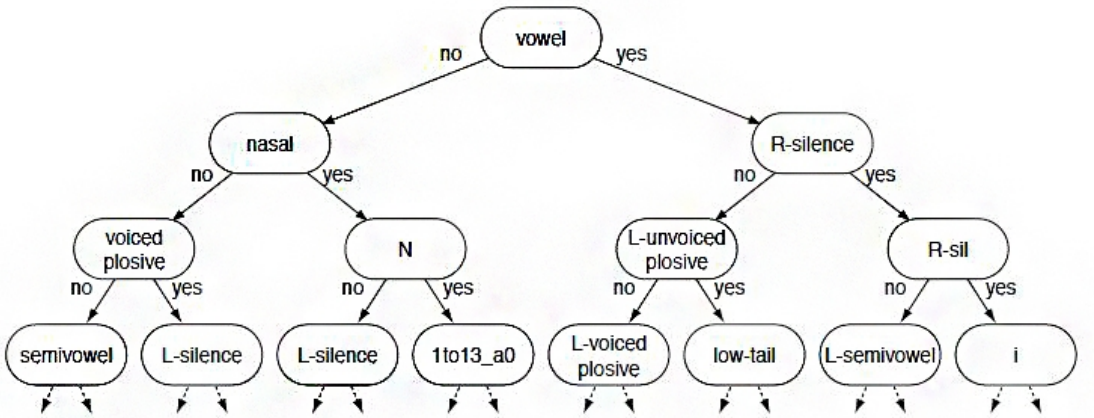

Fig. 2: Example of a phonetic decision tree

In voice building with HTS, further embedded training was applied after context-clustering. This is efficient and necessary to get better parameter estimates. The initial training was conducted with monophone models, which resulted in well-estimated models since enough examples of monophones were available in the training data. Then, context-dependent models were generated, but not all of these context-dependent models could be trained because some of them occurred only once, and some did not occur at all in the training data. Thus, we had very badly trained models. However, this did not matter because all they were used for was for the creation of a tree that did the parameter tying. Context-classes were generated for phones with similar parameters. When the context-dependent phones are clustered, there is enough training data for each model and we get better parameter estimates after re-training. Next, the well-trained context-clustered models are cloned, untied, trained for one iteration and re-tied (cf. Jurafsky and Martin 2009: 382). Because we start from a much better initialisation than in the previous training steps, this training process will result in a better tree. We now have pretty good parameter estimates for synthesis.

2.4 Speech synthesis

In the synthesis part, for a given input text sequence T that is to be synthesized, a set of speech parameters S is generated from the estimated models λ (cf. Zen et al. 2009: 1042):

$$\widehat{S} = \arg\max_{S} \{p(S \mid T, \widehat{\lambda})\}$$

A given word sequence is converted into a context-dependent label sequence. Then the context-dependent HMMs are concatenated according to that label sequence and a sentence HMM is constructed. The speech parameter generation algorithm generates the sequences of spectral and excitation parameters from the sentence HMM. According to these parameters, a

synthesis filter module synthesises a speech waveform (cf. Zen et al. 2007(a): 158). Durations are modelled by multivariate Gaussian distributions and are "determined so as to maximize the output probability of state durations" (Tokuda et al. 2002: 2).

In HTS, synthesis can be conducted with the speech parameter generation tool HMGenS, which applies the EM algorithm, or the software hts_engine (cf. Zen et al. 2007(b): 295). For this dissertation project, sentences were synthesised with hts_engine. Hts_engine basically predicts the state durations by picking the mean of the duration distribution for each state, and, given this state alignment, performs the maximum likelihood parameter generation (MLPG) algorithm. The MLPG algorithm considers the properties of dynamic speech parameters (delta and delta-delta coefficients) and finds the most likely output parameter sequence given the static speech parameter distributions as well as the distributions of the delta and delta-delta coefficients (cf. King 2010: 10)

3 The audio book data

The data that was released for phase 1 of the Blizzard Challenge 2012 was available to build the synthetic voice. This data was provided by Toshiba Research Europe Ltd (cf. http://www.synsig.org/index.php/Blizzard_Challenge_2012) and includes four audio books from librivox.org (http://librivox.org), an internet platform that provides audio books in the public domain, which are not restricted by copyright laws. All audio books on Librivox are recorded by volunteers. The four audio books that were used in this project were obtained from Project Gutenberg. Books from the Gutenberg Project have an accuracy target of 99% for their texts (cf. Prahallad 2010: 30), which means that there are some errors in the text that might have arisen in the transcription process. All audio books were recorded by the same American English speaker, John Greenman, who has a fairly consistent and stable voice, and are based on books written by Mark Twain between 1880-1910 (total running time indicated in parentheses):

1) A Tramp Abroad (15:46:01)
2) Life on the Mississippi (14:47:27)
3) The Adventures of Tom Sawyer (6:46:12)
4) The Man That Corrupted Hadleyburg, and Other Stories (13:04:00)

What has to be considered is that the speaker who records the book is likely to make mistakes and does not provide a one-to-one reproduction of the actual text. Moreover, the speaker adds

information about the source of the book at the beginning and the end of the book, and usually at the beginning of each chapter, e.g. “This is a Librivox recording. All Librivox recordings are from the public domain”. The Blizzard Challenge provided phonetic transcriptions that accounted for these additions, but they did not use a label format that agreed with the format needed for HTK. Therefore, new Festival labels were created from the output that had been retrieved by lightly supervised recognition (cf. Braunschweiler et al.: 2010), and was provided by the Blizzard Challenge. With lightly supervised recognition, the recorded speech had been recognised and the locations of word sequences had been used for automatic time alignment of text and speech. Then, recognised speech and the actual book text had been compared and a confidence measure was calculated, which determines the agreement of the recording and the book text. For the creation of labels, only output with a confidence measure above 90% was extracted. It was manually checked that the difference between the book text and the recording was usually a matter of punctuation or different notation of numbers if the confidence measure ranged between 90 and 100%.

For building the voice, not all data was used due to several reasons. *The Man That Corrupted Hadleyburg, and Other Stories* contained a large amount of German and French sentences that were recorded by a different speaker. Around 1,000 label files that contained foreign utterances had to be deleted. A further problem was that some English transcriptions had corresponding wave files with a German or French translation of the English sentence. To play it safe and avoid a change for the worse in the synthetic voice, it was decided to leave out this audio book completely and keep some of the correct transcriptions for the test set. Altogether, the remaining three audio books contained around 6,000 out of vocabulary words, which was discovered by checking a list of all audio book words against Festival's pronunciation dictionary with a Python script that was specifically written for this task. *A Tramp Abroad* contained a number of German sentences, but they remained within acceptable quantities and were manually removed from the database. Furthermore, sentences that contained German place names or proper names were removed as Festival predicted a pronunciation that was very different from the narrator's pronunciation. Finally, foreign place names and proper names were removed from *Life on the Mississippi*, e.g. words like *Mont Blanc*, for which Festival predicted an English sounding pronunciation.

Out of vocabulary words that were not of foreign origin, were kept in the database, and all in all, the number of out of vocabulary words could be reduced to around 2,000. The main reason for this still high amount is that the book is written in British English, but Festival uses an American pronunciation dictionary for the American English story-telling

voice. Furthermore, Mark Twain uses a number of rare words, e.g. *telescopulist* or *gimcrackery*. However, these words still follow the rules of English pronunciation and it was checked for a selection of out-of vocabulary words that Festival predicted their pronunciation correctly. It is assumed that Festival also predicts the correct pronunciation for the remaining English words which are not in the dictionary. The final amount of data after cleaning up was as follows:

1) A Tramp Abroad (13:51:52)
2) Life on the Mississippi (13:31:36)
3) The Adventures of Tom Sawyer (6:45:36)

For the synthetic voices, not all of this data was used for reasons of time and computational costs. The actual data will be described in section 4.3.

In general, audio books are a good candidates for building synthetic voices if certain conditions are met:

- consistent speaking style
- a large amount of training data is available for a single speaker
- acceptable recording conditions

The speaking style of the speaker sounded quite consistent. Furthermore, the recording environment was quiet throughout and the only disturbance was a faint microphone noise. An analysis of the narrator's speaking rate in terms of syllables per second could have been conducted, but it was decided not to do so. The reason for this was that listening to sound samples from different chapters of the book revealed that his speaking rate seemed to be fairly stable all the time.

In contrast to carefully designed speech databases, such as the CMU ARCTIC database, which contain short out-of-context sentences, audio books contain rich prosody including “intonation contours, pitch accents and phrasing patterns” (Prahallad 2010: 31). This is an extremely good prerequisite for building a natural sounding synthetic voice, but the drawback is that while a manually designed database attempts to capture a very large variety of possible phone contexts, we have to rely on the authors of the book text when dealing with audio books. The number of contexts that are covered by audio book databases cannot be pre-selected and a prior analysis of the data is helpful. In the following sections, the quinphone coverage as well as the prosodic coverage of the audio books will be examined.

3.1 Quinphone coverage of the audio book data

As was pointed out earlier, the amount of quinphone combinations that could logically occur is 41^5. However, in the English language not all of these combinations can be found and 41^5 is not a realistic number the quinphones in the audio book data can be compared to. For that reason, a 1,000,000 word corpus containing data from the same domain as the training data was extracted from Project Gutenberg. This corpus contained books like Charles Dicken's *Oliver Twist*, several Jack London short stories and books by Mark Twain that were not part of the training corpus. A 1,000,000 million word corpus is still not a fair representation of the English language, but to keep computational load to a minimum and still investigate a corpus that is considerably larger than the audio book data, it should suffice for the data analysis. The texts from Project Gutenberg were converted into Festival quinphone label files, and the quinphone types and tokens were counted. The following table gives an overview of the number of words, types and tokens in the audio books and the extracted corpus.

	quinphone		word
	types	tokens	tokens
Tom Sawyer	119,045	264,745	73,512
Tramp Abroad	187,079	483,060	142,764
Mississippi	180,577	456,636	136,284
Tom Sawyer + Tramp Abroad	251,804	747,805	216,276
Tom Sawyer + Mississippi	245,666	721,381	209,796
Tramp Abroad + Mississippi	289,365	939,696	279,048
All audio books	339,864	1,204,441	352,560
Gutenberg corpus	643,492	3,762,335	1,000,749

Table 1: Distribution of quinphone types and tokens over number of words (figures based on the cleaned-up data)

Table 1 and Figure 3 (see next page) illustrate that the number of quinphone types increases with the number of occurring words. *Tramp Abroad* and *Life on the Mississippi* have word counts that are not far apart, and this applies to their quinphone type counts as well. *Tom Sawyer* has approximately half the amount of words of *Tramp Abroad* and *Life on the Mississippi*, and contains around 35% less quinphone types. When we look at two audio books taken together, we can observe a similar situation. *Tom Sawyer + Tramp Abroad* and

Tom Sawyer + Life on the Mississippi have comparable word and quinphone type counts, while *Tramp Abroad + Life on the Mississippi* have a higher amount of words and also a higher amount of quinphone type occurrences.

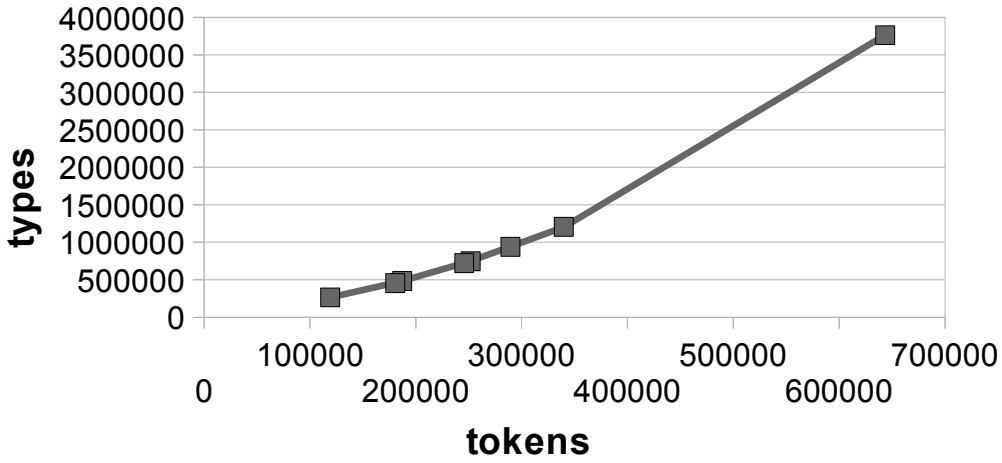

Fig. 3: Distribution of quinphone types vs. tokens

Compared to a larger amount of data, Tom Sawyer covers only 18.46% of the quinphone amount found in the Gutenberg corpus, while *Tramp Abroad* and *Life on the Mississippi* cover 31.06% and 31.10% respectively. Looking at two audio books, *Tom Sawyer + Tramp Abroad* and *Tom Sawyer + Life on the Mississippi* cover 39.31% and 38.18%, and the two largest audio books *Tramp Abroad + Life on the Mississippi* cover 44.97% of the quinphone types in the Gutenberg corpus. Taking all audio books together, 56% of the quinphone amount in the Gutenberg corpus are contained in the training data (see Figure 4). This shows that taking only one audio book for voice building results in a poor quinphone coverage, and that it is a better idea to use more data.

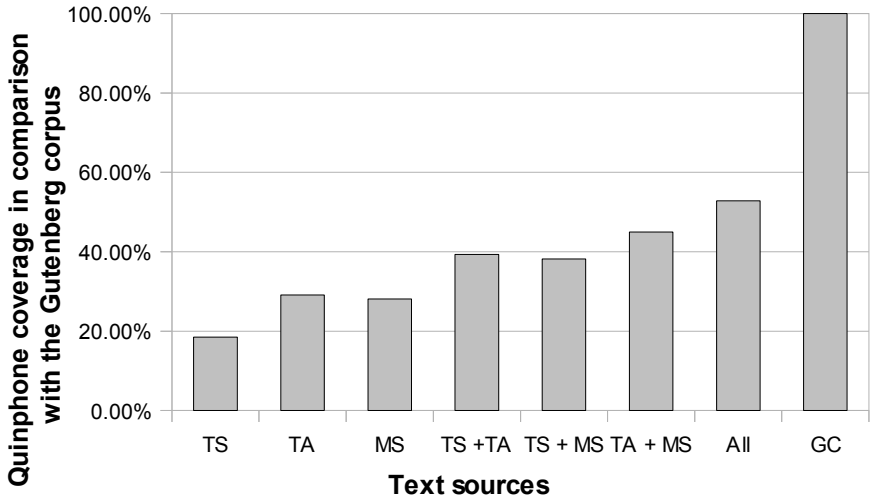

Fig. 4: Quinphone coverage of the audio book data (TS=Tom Sawyer, TA=Tramp Abroad, MS=Life on the Mississippi, GC=Gutenberg Corpus)

However, do we really need such a large variety of quinphone types we find in a 1,000,000 word corpus to build a high quality synthetic voice? Using a large amount of data results in high computational cost. In fact, most quinphone combinations rarely occur more than once, and if they occur more often, they are part of commonly used words. Therefore, it is a better idea to look at the more frequent quinphones and investigate how many of those are covered by the audio book data. For this analysis, the 3,000 most frequent quinphone types were extracted from the Gutenberg corpus and compared to the 3,000 most frequent quinphone types in the audio book data. It was decided to look at the 3,000 most frequent quinphone types because these types occur in a large quantity in the audio book data; each type occurs at least 100 times.

Figure 5 shows that the coverage of the 3,000 most frequent quinphones does not change considerably with a larger amount of data. Reasons for this might be that the Gutenberg corpus is either not a representative sample, or that Mark Twain did not use a wide variety of words in his books. Indeed, an analysis of all words in the three books revealed that at least 25% of the word types occuring in one book occur in one of the other books as well:

Tom Saywer + Tramp Abroad	4,874 (25.28%)
Tom Sawyer + Mississippi	4,793 (25.12%)
Tramp Abroad + Mississippi	6,809 (28.66%)

Table 2: Overlap of word types across 2 audio books

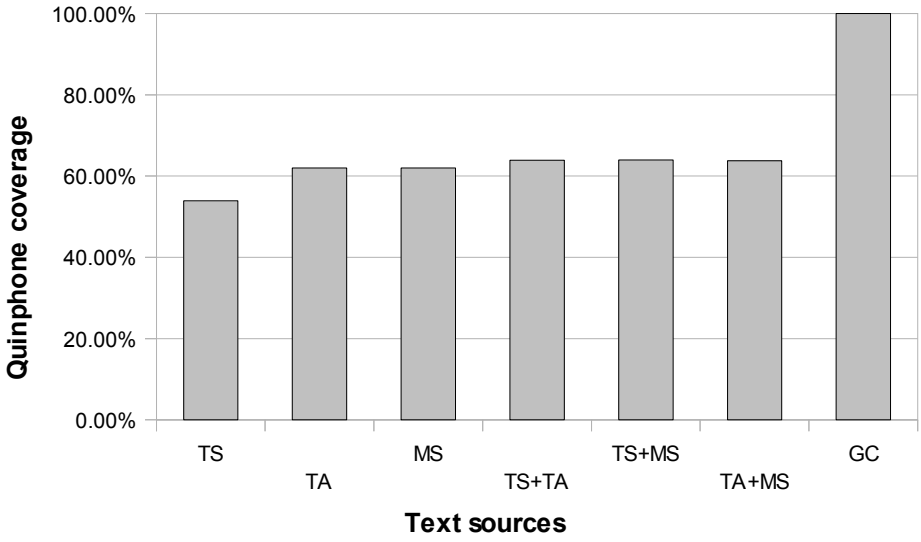

Fig. 5: Quinphone coverage of the 3,000 most frequent quinphones (TS=Tom Sawyer, TA=Tramp Abroad, MS=Life on the Mississippi, GC=Gutenberg Corpus)

The preceding analysis suggests that in order to cover the most frequent quinphones, using just one audio book should suffice. What has to be pointed out, however, is that there are also quinphone types in the audio book data that are not in the Gutenberg Corpus! Looking at all three audio books together, it was found that approximately a third of the quinphones in the audio book data does not occur in the Gutenberg corpus. This applies to taking all quinphone types as well as taking only the 3,000 most frequent quinphone types into consideration. Therefore, only comparing the audio book data to a larger corpus is dangerous! Taking this into account, the quinphone coverage of the audio book data is actually better than was presented in the above diagrams. Nevertheless, there are other things that have to be taken into consideration as well, such as if questions are well modelled, or if a sufficient number of prosodic contexts are covered.

What questions are concerned, they are rare in narrative text, which does not include dialogues between characters. Overall, 156 questions were found in the narrative text in all audio books of which 90 occur in *Tom Sawyer*, 37 in *Tramp Abroad* and 29 in *Life on the Mississippi*. Questions are more frequent when characters talk to each other: 784 questions were found in quoted speech. The reason why a difference is made between narrative text and quoted speech is that the narrator's intonation changes when he mimics characters. Questions in quoted speech are livelier and more emphatic, and often much shorter than in the narrative text. Quinphone coverage in questions is also an important issue. The intonation at the end of questions usually rises and modelling this with quinphones from a declarative sentence would not sound like a question. It has to be made sure that a large variety of quinphones is available to model the end of questions. To investigate this matter, the last 5 quinphones of each question were extracted, and their type and token frequencies were counted. By looking at a number of pitch contours of questions, it was found that usually the last five phones of an utterance have a rising pitch (see Figure 6).

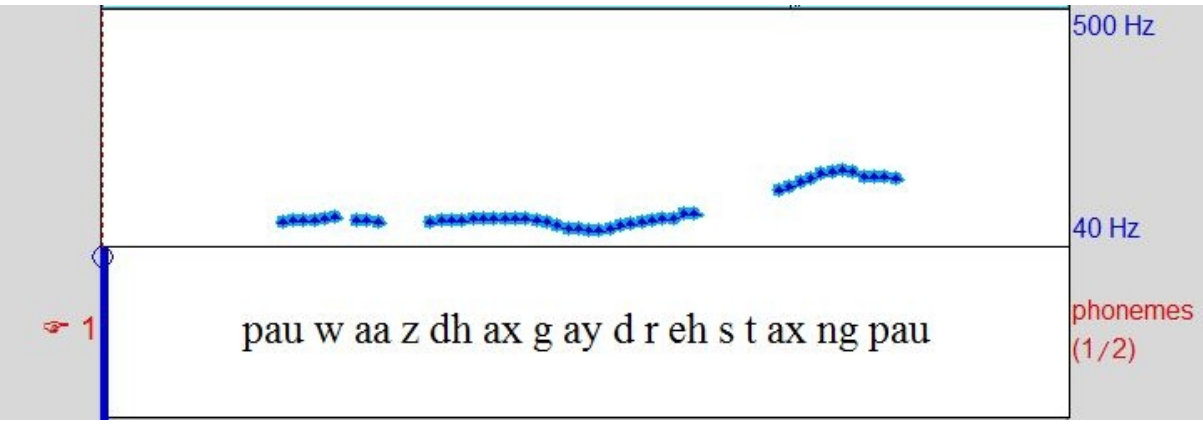

Fig. 6: Pitch contour of a question

Table 3 shows that despite containing the smallest amount of data, Tom Sawyer includes the largest variation of quinphone types at the end of questions. What is noticeable is that there is hardly any overlap of quinphone types: Overall, only 105 quinphone types occur in more than one audio book. However, this is no surprise if we look at the large number of different quinphone types in the whole database compared to the rather small number of quinphone types here. If we want to achieve the best possible modelling of questions, the questions from all audio books should be included in the training database.

	types	tokens
Tom Sawyer	1,299	1,917
Tramp Abroad	694	980
Mississippi	790	1,007
All audio books	2,378	3,904

Table 3: Distribution of quinphone types and tokens at the end of questions in quoted speech

3.2 Prosodic variation

The intonation of a speaker has an effect on the fundamental frequency (f0) of his speech. Since the audio books contain a variety of intonation styles such as character mimicking, it can be assumed that a wide range of f0 values is covered. When the speaker mimics characters, he often uses a higher pitched voice, but sometimes applies a dark voice as well. Therefore, the f0 range will probably be larger for quoted speech (character mimicking) than for narrative speech (the speech style which is employed when no characters are mimicked). For this reason, narrative speech and quoted speech will be analysed separately. Histograms were computed to determine and compare the pitch ranges of narrative and quoted speech. The f0 values were put into frequency bins, each of which has a frequency range of 10 Hz, and plotted on a logarithmic scale. A logarithmic frequency distribution is more Gaussian-like and, moreover, the logarithmic scale approximates the sense of human hearing (cf. Jurafsky and Martin 2009: 269). The notation is given in Hertz values to facilitate understanding. Because the average f0 did not differ significantly between the three audio books for both narrative and quoted speech, the results are summarised in one histogram respectively.

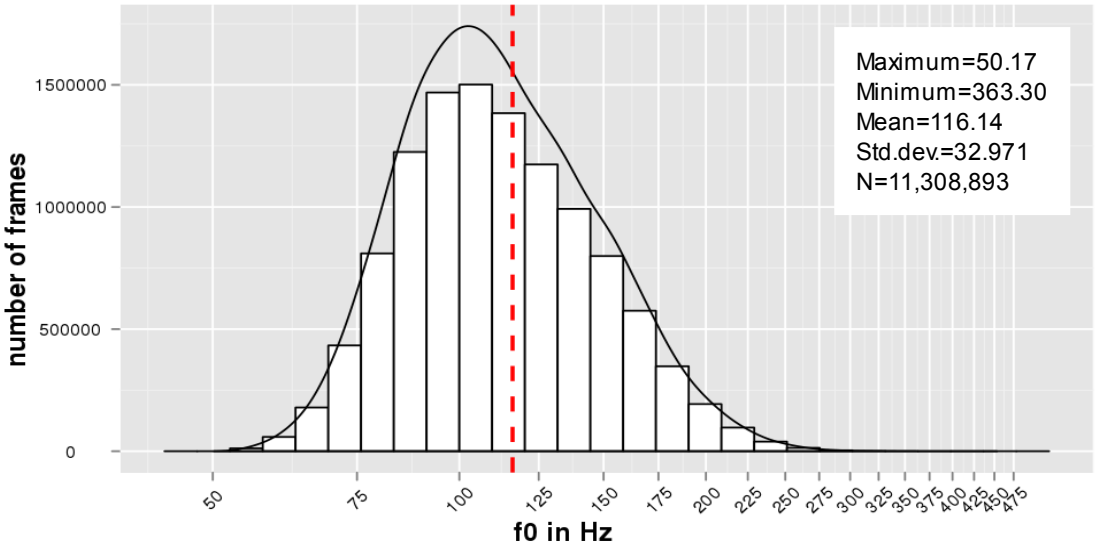

Fig. 7: Distribution of f0 in narrative speech on a logarithmic axis

Figure 7 illustrates that most f0 values lie between 80 and 140 Hz. There are some frames in the higher frequency range above 200 Hz, which usually do not make up a whole utterance, but arise when the speaker tries to convey excitement and his voice becomes very high and cracks. In addition to the distribution of framewise f0, the average f0 values for each utterance were analysed. This eliminates outliers in the upper frequency range and serves as the evidence that whole utterances with a very high overall frequency do not occur in narrative speech.

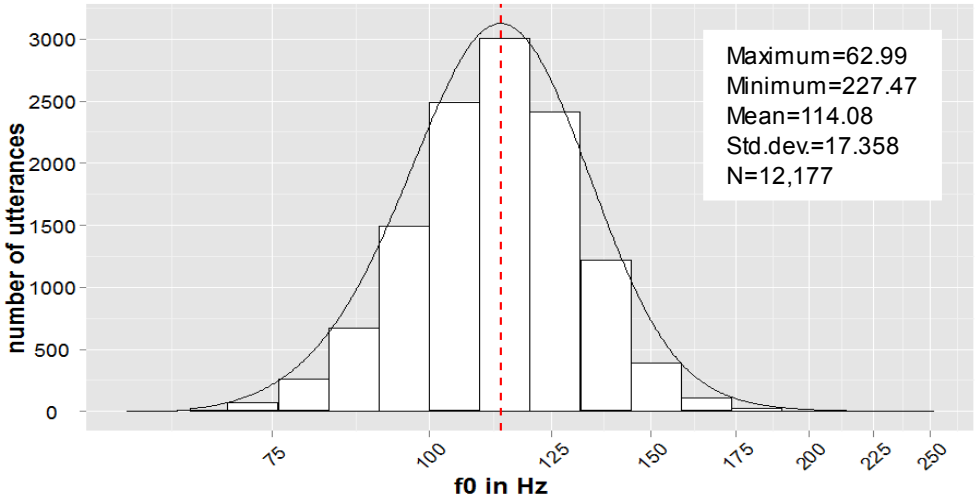

Fig. 8: Distribution of average f0 per utterance in narrative speech on a logarithmic axis

As Figure 8 shows, the highest average f0 for a whole utterance is around 227 Hz. However, average frequency values in this range are rare. The distribution follows a normal distribution and stands in contrast to the distribution of framewise f0 in Figure 7, which is right-skewed. This underlines the argument that the distribution of high and low f0 values within an utterance is fairly balanced if we look at average f0 per utterance.

As it was expected, the frequency range is wider for quoted speech, namely from 49 to 475 Hz (see Figure 9). Most f0 values lie between 80 and 180 Hz, which is a wider range compared to narrative speech. The mean, 141.26 in contrast to 116.14 in framewise f0 for narrative speech, is considerably higher as well. Moreover, in comparison to narrative speech, there is a considerable amount of f0 frames above 200 Hz.

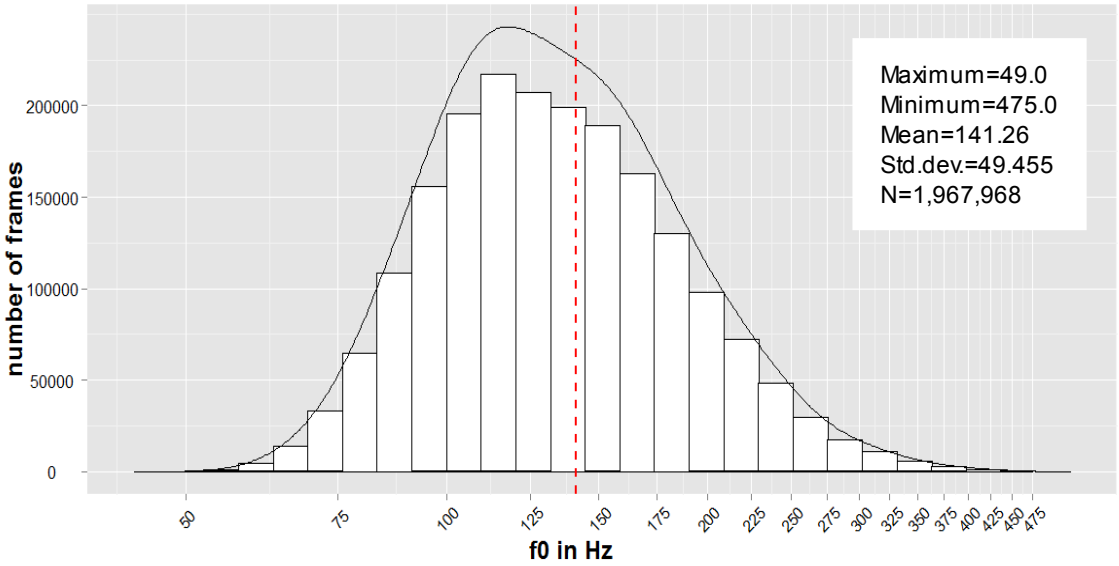

Fig. 9: Distribution of framewise f0 in quoted speech on a logarithmic axis

Figure 10 shows that this distribution does not differ considerably when we look at the average f0 value for whole utterances: both distributions are right-skewed, which suggests that the distribution of high and low pitch is not balanced within an utterance, but that there is actually a substantial number of utterances that are spoken with a high pitch throughout.

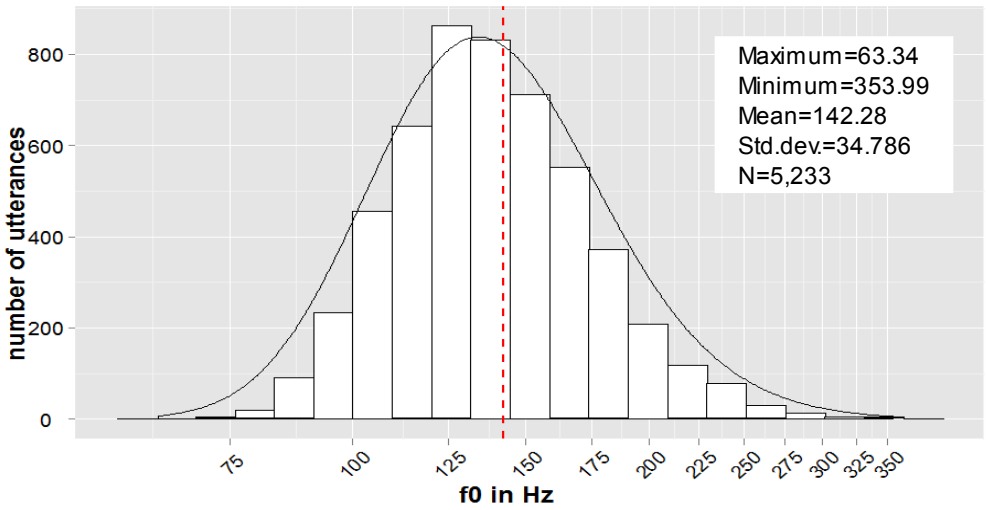

Fig. 10: Distribution of average f0 per utterance in quoted speech on a logarithmic axis

These differences in f0 between narrative speech and quoted speech suggest the need to model these speaking styles differently. A method, which is proposed in section 4.2.2, describes the clustering of f0 according to frequency bins and it will be argued that this method will improve the quality of synthetic speech. The next chapter will describe the techniques that were applied in voice building.

4 Techniques and methods

As the data analysis has shown, the speech data is quite varied: the audio books contain speech in narrative style as well as quoted speech with higher pitch, and there is a large prosodic variation, especially in quoted speech. Before training was conducted, the speech data was divided into separate context-classes according to their f0 values, which will be explained in detail in this chapter. Moreover, noise was removed from the audio files to improve the quality of the synthetic voice.

4.1 Noise removal

Due to the fact that the audio books were not recorded in a studio, the recordings contained a small amount of background noise. As a result, the synthetic speech sounded buzzy and the background noise was audible. The noise was a constant microphone noise which did not change considerably throughout the recording sessions. Therefore, a good solution was to use a multi-band digital noise gate, which was applied with Audacity, an open-source software for recording and editing sounds (http://audacity.sourceforge.net).

Audacity's noise removal algorithm uses Fourier analysis, which is a technique that decomposes any complex waveform into sine waves of different frequencies (cf. Jurafsky and Martin 2009: 332). A noise profile is provided by selecting a part of audio with only noise and Fourier analysis finds the frequency bands that make up the noise spectrum. This noise profile serves as the sample for the background noise in the wave files the spectral noise gate is applied to. The noise floor in each of the frequency bands of the sample is calculated and used as a threshold for a bank of noise gates.

When the noise gate is applied to noisy speech, the algorithm decomposes the speech into its frequency bands, and any pure tone which is substantially louder than the pre-defined threshold will be reduced in loudness. This removes noise not only from silence, but also from speech signals. As a result, noise removal might have a negative effect on the speech signal, but a careful adjustment of the threshold and listening to the noise-free speech revealed that it is a very convenient method.

Noise removal was applied for each audio book separately because careful listening to the recordings revealed that the background noise was at least slightly different in one audio book, namely *Life on the Mississippi*. It might have been a good idea to analyse the energy in every single chapter and see whether there is a considerable difference as it can be assumed that the speaker made a pause between chapters and changed his recording conditions. However, building a small voice from noise-free data revealed that this was not necessary, and noise removal was successful without having a negative impact on the speech itself.

4.2 Features added in voice building

4.2.1 Modelling of quoted and non-quoted speech

Since the speaker employs different speaking styles when he mimics characters and usually speaks with a higher pitched voice, it was decided to build different acoustic models for quoted speech, i.e. character mimicking, and narrative speech (henceforth called *non-quoted speech*). It is expected that modelling these features will improve the synthetic output over a voice which is built from quoted and non-quoted material without any modelling of these speaking styles.

For the identification of quoted and non-quoted text, a Python programme was written that extracted text between quotation marks from the book transcription, which was provided by the Blizzard Challenge, and wrote quoted and non-quoted text together with the corresponding label name into separate files. The text recognised by the lightly supervised approach could not be used for this task since it did not contain any punctuation marks. Unfortunately, quotation marks were not used consistently so that sometimes a non-quote was marked as a quote when a quotation mark was missing or vice versa. Therefore, all extracted data was manually checked for errors. To facilitate the task, an utterance was only extracted if it was either quoted text or non-quoted text, but not a mixture of both, so sentences like, “'Tom', she screamed.” were discarded. However, most utterances were either purely quoted or non-quoted text.

For each non-quoted utterance, a feature label was added at the end of each full-context model in the label files, for instance:

x^x-pau+ao=l@x_x/A:0_0_0/[...]**Q:_0**

Each full-context model got the label **Q:_0** attached to its end. The label itself was defined in a so-called question file, which contained questions relating to each of the features that were modelled, e.g. phonetic context, number of syllables, part-of-speech-tags. The question that

asked for non-quoted material looked as follows:

QS "No_quotes_in_utterance" {*/Q:_0*}

Since context-clustering is conducted using binary decision trees, extra labels for quoted speech are not necessary. If the question " No_quotes_in_utterance" is asked and answered with *yes*, all non-quoted material is clustered at the leaf of the *yes* branch of the tree, and if the question is answered with *no*, every full-context model that does not have the label **Q:_0** at its end is clustered at the leaf of the *no* branch of the tree.

4.2.2 Modelling of f0

As was demonstrated in chapter 3.2, the audio books contain a large prosodic variation, especially in quoted speech. It is assumed that creating models for different frequency ranges will improve the synthetic voice considerably when training data from similar frequency ranges is pooled together. Furthermore, if modelling higher or lower pitched voice for quoted speech is successful, the feature labels that model these speaking styles can be attached to the full-context models used at synthesis time so that the synthetic voice can speak in higher or lower pitched voice. How this is accomplished will be explained in section 4.5.

For f0 modelling, the f0 range was divided into 3 parts: lower pitched speech, average pitched speech and higher pitched speech. Figure 11 shows that the mean of log f0 in quoted speech is approximately 142 Hz. The standard deviation is 34.8 and it can be assumed that the f0 values that are farer away from the mean than the standard deviation sound substantially different.

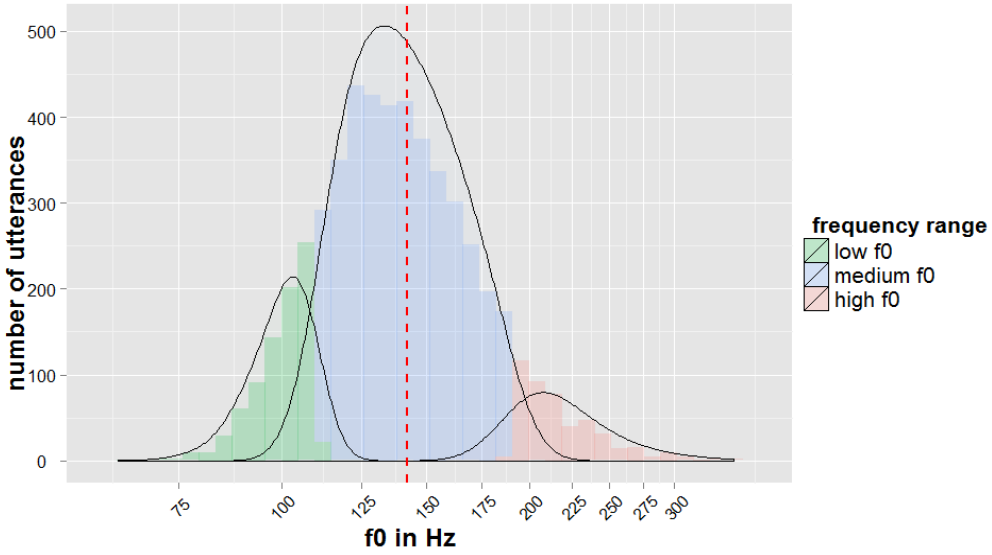

Fig. 11: Frequency ranges in quoted speech on a logarithmic axis

Listening to several audio samples revealed that utterances with an average f0 value above 190 Hz and below 110Hz sound considerably different from utterances spoken with average f0. These higher f0 values are farer away from the mean than the standard deviation, and differ to a large extent from the average, and the lower f0 values are nearly as far away from the mean as the standard deviation. Therefore, it was decided to model a low frequency range from the lowest value 63 Hz to 109 Hz, an average frequency range from 110 Hz to 189 Hz and a high frequency range from 190 Hz to 354 Hz. It is of course impossible to argue that an utterance spoken with an average frequency of 189 Hz is different from an utterance with an average frequency of 190 Hz, but the average f0 values of each individual frequency bin should give a pretty good representation of each frequency range. Moreover, it was decided to use only three frequency ranges for acoustic modelling in order to ensure that the largest possible amount of training data is available for each of the ranges. This is important when the features labels are added at synthesis time because we want to synthesise our sentences from well-trained data.

The same procedure was conducted for non-quoted speech. Although the f0 range for non-quoted speech is not that diverse, it might still result in a better synthetic voice, if similar frequency ranges are clustered together. Listening to the audio book data also revealed that the speaker's voice can sound very different in lower frequency ranges. As was reported before, and can be seen in Figure 12, the tails of the frequency distribution range over different frequencies than it was the case in quoted speech.

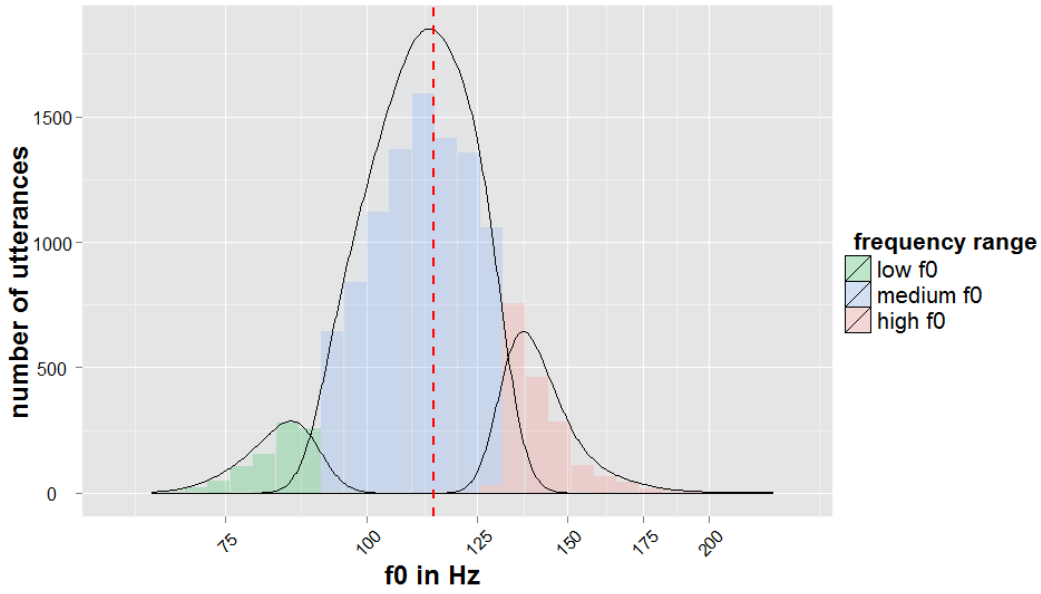

Fig. 12: Frequency ranges in non-quoted speech on a logarithmic axis

Therefore, the frequency ranges that were determined for acoustic modelling differ as well. Again, the tails of the distribution were cut off, this time below 90 Hz and above 130 Hz. These values are more than one standard deviation away from the mean (mean=114.08, std.dev.=15.36), and listening to the audio in these frequency ranges showed that the utterances sound substantially lower or higher pitched. The addition of feature labels and creation of questions was modelled in the same way as described in section 4.1.1, with the exception that questions were needed for each of the frequency ranges:

QS "Quote_low_F0" {*/S:_0*}

QS "Quote_medium_F0" {*/S:_1*}

QS "Quote_high_F0" {*/S:_2*}

QS "Non-quote_low_F0" {*/S:_3*}

QS "Non-quote_medium_F0" {*/S:_4*}

QS "Non-quote_high_F0" {*/S:_5*}

4.2.3 Modelling of questions and exclamations

Two questions were used to model questions in quoted and non-quoted speech separately because due to the prosodic difference between both speaking styles, it is assumed that different parameters should be used to model these contexts. These two questions are:

QS "Quote_with_a_question" {*/R:_0}[2]

QS "Non-quote_with_a_question" {*/R:_1}

The feature 'R' with the value '0' is attached to all labels within questions that occur in quoted material, and the feature 'R' with the value '1' is attached to all labels within questions that occur in non-quoted material. In the case that questions in quoted and non-quoted material have similar parameters, a third question was added, which would cluster quoted and non-quoted questions together if their parameter values are similar:

QS “Quote_or_non-quote_with_a_question” {*/R:_0,*/R:_1}

The higher level structure of the model is determined manually, but which of the questions is asked in the end will be decided by the data. In this case, it is assumed that different parameter values should be used to model different contexts, the possible contexts are

2 No wildcards are needed at the end of these feature labels because they occur at the end of the full-context labels in the corresponding label files

defined, and the learning algorithm uses the data to determine what the actual parameter values are.

It should be noted that only utterances that consisted of a single question were marked as questions. Utterances that contained a declarative sentence and a question were discarded for reasons of simplicity. A Python programme was written that detected the questions by looking at those utterances from the book text which contained a question mark and a preceding utterance, but no preceding or succeeding declarative sentence or exclamation. Their corresponding label names were written into a list and features were added to the full-context labels in each of the label files.

Exclamations were extracted the same way, but were only modelled for quoted speech. The reason for this is that exclamations in non-quoted speech do often not sound different from declarative sentences. Exclamations in quoted speech are generally spoken in a very expressive and lively manner. The audio book data contained 860 exclamations in quoted speech, and it was decided that it is worth making the attempt to achieve a better modelling of these exclamations.

4.3 The voices

It was decided to build two voices from the same large database of utterances. In order to save computational time, but still achieve to get a good quality synthetic voice, not all audio books were used in the training database. *Tom Sawyer* and *A Tramp Abroad* were taken for training. These audio books contain the 2nd largest number of quinphone types, which should serve as a compromise between taking a higher amount of data, which would imply a higher computational cost, and maybe covering too few data types. Each voice was trained on around 11,000 utterances.

One of the voices was modelled with the 11 additional features that were described above. The other voice was built without any modifications and serves as the baseline the modified voice is compared to in the evaluation. The reasons and assumptions behind this choice will be pointed out in detail in the evaluation section of this paper.

4.4 Analysis of decision trees

The decision trees for log f0, mel-cepstral coefficients (MCEP) and duration were analysed with respect to the newly added questions. It is expected that trees for these different features will prefer different subsets of the common question set. To illustrate the preference of questions in the trees, a so-called *dominance score* was calculated (cf. Chomphan 2011), which is defined as the reciprocal of the distance between the root node and the question

node. A node is more important if it is near the root of the tree because the first questions separate more data than later ones. Thus, distance to the root node is correlated with dominance, but it is not a positive correlation; the bigger the distance, the less the dominance - the smaller the distance, the bigger the dominance. This can be called an *inverse relationship*. The reciprocal works like this: as x increases, 1/x decreases, and as x decreases 1/x increases (as long as x is positive). Therefore, if we let x='the distance to the root node' then 1/x behaves like we want 'dominance' to.

For MCEPs and log f0, the dominance scores of the 11 additional questions were computed for each of the 5 emitting states of the HMM, and their average score was calculated. Duration was modelled with a single tree and set of distributions at the phone level. With 5 states, a 5-dimensional distribution was used. Therefore, the dominance scores for phone durations reflect the actual scores that were computed from a single tree. The following diagram gives an overview of the questions and their corresponding dominance scores:

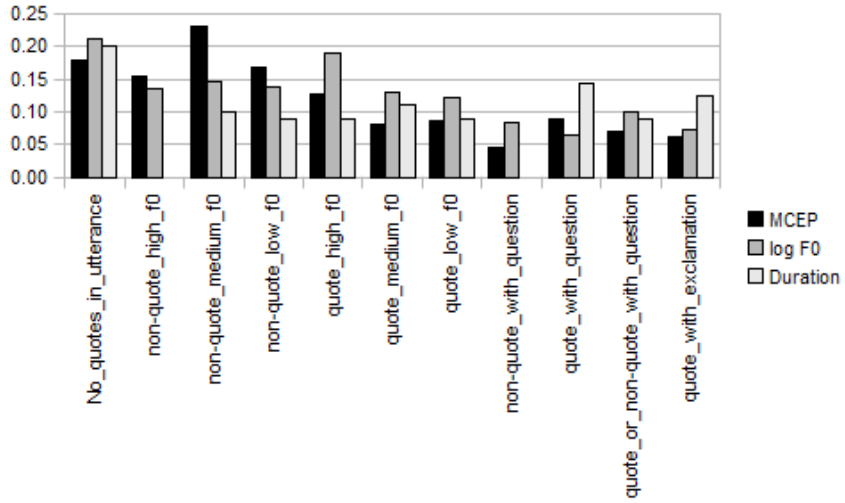

Fig. 13: Dominance scores of questions in context-custering

Figure 13 shows that the dominance scores of questions differ between the trees, but that there are also some comparable tendencies. The first thing to note is that in the MCEP and log f0 trees all questions were regarded as relevant, while all but two questions were asked in the phone duration tree (*non-quote_high_f0* and *non-quote_with_question*). With a dominance score of 0.23, the question *non-quote_medium_f0* occupies the highest node under the 11 questions in the MCEP tree. This means that, among the 11 questions, this question separates the data into the two largest sub-sets. *No_quotes_in_utterance* is the second most important question, with a dominance score of 0.18. This question is nearest to the root node in the log

f0 and phone duration trees. This emphasizes the need for this question and shows that there is a substantial difference between non-quoted and quoted speech with regard to energy and f0, but also phone duration.

In both log f0 and MCEP trees, questions relating to the f0 values in non-quoted speech received a relatively high dominance score as well, which suggests that there is a high amount of data that shares the features of non-quoted speech with high, medium and low f0 respectively. What is noticeable is that the question *quote_high_f0* received the second highest dominance score in the log f0 tree (0.19). This was expected since there is a considerable amount of utterances in the quoted speech data that is spoken with a higher pitch. Clustering the data into parameters that share the features of quoted speech with higher f0 is therefore a reasonable thing to do.

While the dominance scores in the phone duration tree are higher for the questions *quote_with_question* and *quote_with_exclamation*, this does not apply to the MCEP and log f0 trees. Listening to the audio files revealed that the phone durations of exclamations in quoted speech seem to be shorter than in narrative speech as these are usually spoken in a very lively and hasty manner, which could be the reason for the higher dominance score in the duration tree for *quote_with_exclamation*. A reason for the lower dominance scores in MCEP and log f0 trees might be that in comparison to declarative sentences, questions and exclamations are rare, and can therefore not be split into large sub-sets and have to be clustered further down in the trees.

All in all, it can be assumed that the additional features are well-modelled. All questions were asked at the upper nodes of the trees, and adding their corresponding features at synthesis time will probably be successful. How the attachment of these features was accomplished, will be the topic of the following sub-section.

4.5 Features in synthesis

For synthesis, sentences from the held-out audio book *The Man That Corrupted Hadleyburg, and Other Stories* were manually selected, extracted from the book text and converted into full-context label files with Festival. For each domain to be synthesised, corresponding sentences were chosen. That is to say, quoted speech was synthesised from text that was spoken by characters, non-quoted speech was synthesised from narrative text, and questions and exclamations were synthesised from corresponding sentences in the book text. Because no feature was defined for quoted speech in context-clustering, features relating to the f0 ranges in quoted speech were added here.

It was decided manually, where specific features should be added. The addition of features worked the same way as in the preparation of label files before training. A shell script was written that looped over the full-context labels in each label file and added the required feature label to the end of each label. It should be noted that the more features are added, the fewer parameters to synthesise a sentence might be available. For example, if we want to synthesise quoted speech with high f0, and add a label for this at synthesis time, the resulting speech output might sound quite good. But if we add a further label, for instance, if we would like to synthesise a question with a higher pitched voice, the speech might sound buzzier. This is the case because the pool of parameters that can be used at synthesis time becomes more restricted when more features are added.

4.6 Automation of feature modelling in synthesis

For the use as an actual applet in an e-book reader, the process of adding features at synthesis time should be automated. For the listener, it would be very uncomfortable if he had to tell the reader when to change his pronunciation and such an implementation would be quite impractical. One suggestion to solve this problem is to apply classification and regression trees (CART) (cf. Bennett and Black: 2005). CART trees are decision trees and work like it was described in section 2.3. While classification trees are used to predict the value of a categorical variable, regression trees predict the value of a continuous variable. For the features that are added at synthesis time, a classification tree can be used.

Classification trees can be learnt from data, but the pre-condition for this is that the predictors, the predictee and the questions that are asked about the predictors are pre-defined. The predictee might be one of the features we want to add at synthesis time, and the predictors are the characters of linguistic or prosodic items that help to classify a sequence of words or letters as this particular feature. For instance, if we would like to detect a quoted utterance in a book text and synthesise this utterance as quoted speech with a different pitch, example questions could include, “Is the utterance surrounded by quotation marks?” or “Is the utterance preceded by a [insert a word that signifies the act of talking]?”. As it was the case in context-clustering, the training algorithm will choose the questions that partition the data into the most consistent sub-sets and will decide where to put the questions in the tree.

When these classification trees are implemented into the e-book reader, they will automatically decide, which sentence has to be uttered in a particular way. However, more research is necessary to prove this argument and see in how far this approach is technically feasible, and, moreover, if listeners approve this kind of implementation.

5 Evaluation

The purpose of the evaluation was to find out if listeners prefer a voice that sticks to the same speaking style throughout the narration of an audio book, or if a voice that can apply different speaking styles is equally acceptable. For this reason, a baseline voice built from around 11,000 utterances was compared against a modified voice that included 11 additional features and was built from the same database as it was described in chapter 4.3.

Listening to the voices revealed that the modelling of most features was successful. The modified voice is able to talk with lower and higher pitched speech, and there is an audible rise of intonation at the end of most questions. Furthermore, non-quoted speech has a good quality and nice prosody. The prosody of the baseline voice is quite good and natural sounding as well, but careful listening to the modified voice revealed that its prosody is slightly richer and sounds more natural. However, when the voice mimics characters and talks with a different pitch, it often sounds buzzier than non-quoted speech of both the modified voice and the baseline. Moreover, questions are not perfect: their intonation is not always smooth and stable and they sometimes sound quite buzzy as well. Exclamations are not well-modelled at all and not identifiable as such. This is due to the fact that not enough training data was available for high and low pitch ranges, questions and exclamations. One way to improve this is to add more of this material to the training data, preferably by the same speaker, or, if this is not available, data from a different speaker and adapt a transform. Because of the buzziness of quoted speech and questions, subjects might prefer the baseline voice. If this is the case will be answered later.

5.1 Experimental design

The evaluation was implemented as an online survey, which was set up as follows. The aim of the survey was to assess the naturalness and liveliness of the two voices as well as their commercial viability, i.e. if people would buy the voices. Although it is always good to test the intelligibility of a voice, a conscious decision was taken not to do this for several reasons. First, semantically unpredictable sentences (SUS) (cf. Benoît et al. 1996) were not available in natural speech. These sentences are syntactically correct but meaningless, and do not contain any semantic contextual cues. This is important for testing the intelligibility of a speech synthesis system because it ensures that listeners cannot guess the following words in a sentence when they write down what they understood, and have to rely on their ears only. Because listeners also make mistakes when they listen to natural speech, it is crucial to test both natural and synthesised speech to determine the significance of the speech synthesiser's

degree of intelligibility. For an intelligibility test, semantically predictable sentences would have had to be used. This is problematic because the degree of correctness of a sentence can depend on the linguistic competence of a speaker. Some speakers might be better than others at guessing a particular word they did not understand when the semantic context is provided. Another reason for excluding the intelligibility test is that the e-books which contain the test sentences are freely available on the internet and can be found with a simple web search. It is not claimed that subjects are inclined to cheat, but the possibility is there, and usually subjects want to perform well, even if the test is anonymous. The results of an intelligibility test are therefore not reliable enough.

Each of the three tests contained three sub-sections, which were the same for all tests: non-quoted speech, quoted speech and questions were assessed separately. Exclamations were excluded from the test since they could not be identified as such. Furthermore, only higher pitched speech was included in the evaluation of quoted speech because it had a better quality and was easier to distinguish from non-quoted speech than lower pitched speech. Questions were synthesised as non-quoted speech because this allowed a reasonable comparison to the baseline.

For each sub-section, a total number of 40 sentences were synthesised. For each of the voices, the same 20 sentences were taken so that a proper comparison can be made between the two voices. From the pool of 40 utterances, 20 random sentences were played to the subjects. It was ensured that the subjects did not listen to the same sentence twice. If a sentence from the baseline voice was selected, it was controlled that the same sentence was not re-played with the modified voice. Although the three tests contained the same sub-sections, the synthesised sentences differed in each corresponding sub-section. This way, a larger amount of data can be evaluated, the evaluation has more variety and becomes more interesting for the subjects. How the individual tests looked like will be described in detail in the sections for each test.

5.2 Participants

The number of participants was the same for all tested domains. Altogether, 39 participants took part in the evaluation, of which 32 had no experience with speech technology and never or a few times ever listened to a synthetic voice. 6 participants had experience with speech technology and listened to synthetic voices regularly. It was ensured that only native English speakers participated. Before the participants entered the survey, they were asked if they are native English speakers or not, and if they answered the question with *no*, they were not able

to enter. Moreover, the participants were asked the same question at the end of the study again, and one participant who finally admitted to be a non-native speaker was excluded from the statistical analysis. Therefore, the statistical analysis reflects the results of only 38 participants. It was assumed that including non-native speakers in the evaluation is problematic. Depending on their level of language proficiency, some speakers understand the content of the sentences better than others and this might influence their rating.

5.3 Evaluation of naturalness

5.3.1 Methodology

In the evaluation of naturalness, the subjects listened to one voice per page, and rated it on the basis of five categories: completely unnatural, most unnatural, equally natural and unnatural, mostly natural, completely natural. Only one category could be selected. For the statistical evaluation, these categories were transformed into opinion scores so that most unnatural received a score of 1 and most natural a score of 5. For each voice, the mean opinion score (MOS) was calculated by adding up the scores and dividing them by the number of sentences that were played to each subject, namely 10 per voice. Descriptive statistics and a one-way repeated measures ANOVA to analyse the within-subjects and between-subjects effects were carried out by means of SPSS (Version 17.0). The tests were the same for non-quoted speech, quoted speech and questions.

5.3.2 Results

The hypothesis was that there is no significant difference in naturalness between the baseline voice and the modified voice in non-quoted speech. Listening to the voices before the evaluation was carried out did only reveal subtle differences. In fact, both voices sounded very similar and especially for inexperienced subjects it might be difficult to perceive any difference at all. For experienced subjects, the richer prosody is possibly noticeable. There may also be a difference in MOS between inexperienced and experienced subjects because their perception of naturalness might differ.

In addition to that, it was assumed that there is a significant difference in naturalness when it comes to quoted speech. The hypothesis is that quoted speech is regarded as less natural sounding when spoken by the higher pitched modified voice because it sounds buzzier. Moreover, although character mimicking often involves speaking with a higher pitch, the subjects might not appreciate the way this is accomplished by the modified voice because it sometimes sounds rather creaky. This is expected to be the case for both experienced and inexperienced subjects.

What questions are concerned, the modified voice performed quite well and it is expected that the subjects will agree with that and prefer the modified voice in terms of naturalness. However, it might also be the case that they regard the baseline voice as more natural, although it lacks a rising pitch at the end of questions, because the modified voice does not sound smooth and stable throughout as it sometimes selects parameters from non-questions due to an insufficient amount of training data for questions. If it is the case that some subjects prefer the baseline voice and others prefer the modified voice in terms of naturalness, then there might be no significant difference between the two voices.

5.3.2.1 Descriptive statistics

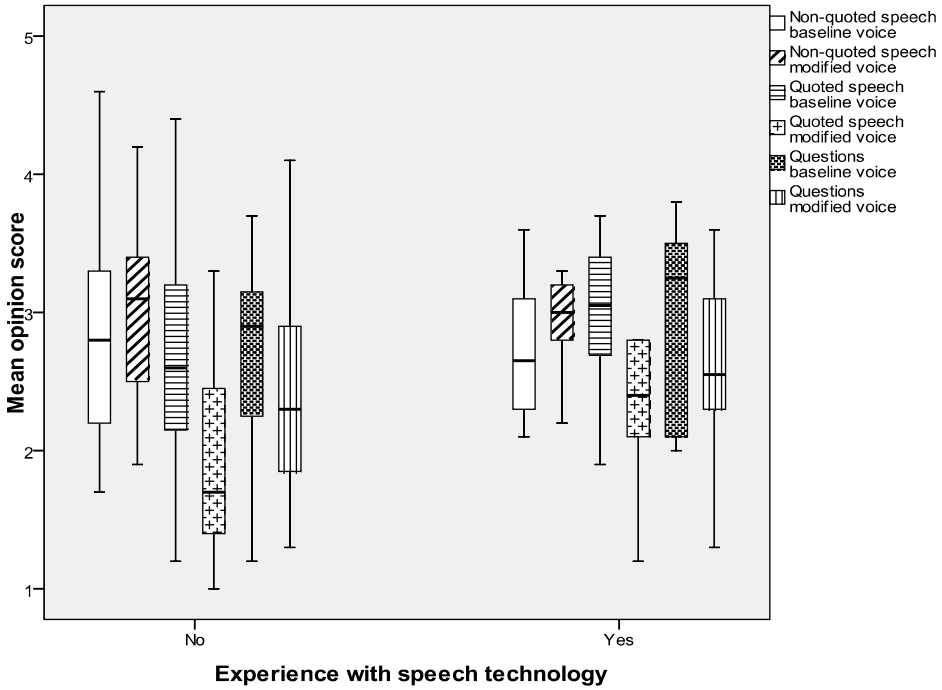

Fig. 14: Variation of MOS for naturalness

The descriptive statistics in Figure 14 show that in matters of non-quoted speech, the MOS of the modified voice is slightly higher for both experienced and inexperienced listeners. The MOS of the baseline voice is 2.85 and 2.73, and the MOS of the modified voice is 2.98 and 2.91 for inexperienced and experienced listeners respectively.[3] Moreover, the lower bound of the MOS of the modified voice is higher for experienced listeners than for inexperienced listeners, which suggests that some experienced listeners heard a difference between the two voices, and a look at the individual results confirms this.

3 Note that the horizontal lines in the boxplots represent the median, although in most cases that were analysed here the values for mean and median are not far apart.

If we look at the descriptive statistics for quoted speech, it is noticeable that the baseline voice was perceived as substantially more natural than the modified voice. While the MOS of the baseline voice is 2.66 for inexperienced and 2.97 for experienced subjects, the MOS of the modified voice is only 1.94 for inexperienced and 2.28 for experienced listeners. What is interesting is that the MOS for the baseline voice differ to some extent from the MOS in non-quoted speech, although the quality of the voice did not change at all. Inexperienced subjects rated the baseline voice in quoted speech slightly lower, while experienced subjects rated it slightly higher. A look at the individual data points showed that this applied to around 2/3 of the inexperienced subjects and also 2/3 of the experienced subjects. A reason for this might be that the inexperienced subjects still regarded the baseline voice in quoted speech as somewhat natural, but in relation to the fact that it is supposed to be used in character mimicking, they might have expected a different voice quality, however, not the one applied by the modified voice either. For the experienced subjects, it might have been the case that they wanted to point out the difference between the modified voice and the baseline voice in quoted speech, and therefore set the scale of naturalness slightly higher for the baseline voice than before.

Another point that is worth mentioning is that the lower bound of MOS of the baseline voice is higher for experienced subjects than for inexperienced subjects in quoted speech, and, even more dominant, the experienced subjects perceived quoted speech spoken by the modified voice as much more natural than the inexperienced listeners. This suggests that experienced listeners have a different perception of naturalness when they listen to synthetic voices. When inexperienced subjects hear a synthetic voice, they might have higher expectations as they are not familiar with the limitations of synthetic speech and therefore are more critical in their judgements. On the other hand, experienced listeners know the state-of-the-art and do not expect a synthetic voice to sound exactly like natural speech, which results in less critical judgements.

Looking at the descriptive statistics for questions, it is obvious that the baseline voice was perceived as more natural again. The baseline voice has a MOS of 2.7 and 3.0 for inexperienced and experienced subjects, and the modified voice has a MOS of 2.4 and 2.6 for inexperienced and experienced subjects respectively. Again, the experienced listeners rated both voices slightly higher than the inexperienced listeners, but both groups agree that the lack of rising intonation makes the voice more natural sounding, probably because the baseline voice sounds considerably more stable. Although it looks like some of the experienced listeners rated the modified voice slightly higher than the baseline (see Figure

14), a closer look at the individual data points showed that this is actually not the case. All in all, there seems to be a general agreement of MOS between the subjects and there are no outliers. If the differences between the voices and the subject groups are significant, will be shown by the following inferential statistics.

5.3.2.2 Inferential statistics

A one-way repeated measures ANOVA was applied to investigate if the MOS of the baseline voice and the modified voice differs within subjects. Moreover, it was analysed whether experience with speech technology has a significant effect on the judgement of naturalness.

Levene's test was not significant, i.e. the variances between the two groups of voices are equal, which is a pre-condition for a parametric test like ANOVA, and the significance values can be trusted. This applies to all statistical tests that were conducted. The results of the sphericity assumed within-subjects tests in Table 4 show that for the voices in non-quoted speech, the mean square, which represents the amount of variation due to the experimental manipulation (cf. Field 2005: 322), is quite low (MS=0.255). So is the F-ratio, which indicates that the experimental variation has been unsuccessful (F=2.716), which is confirmed by the non-significant p-value ($p > 0.05$). This means that there is no significant difference between the MOS of the baseline voice and the modified voice in non-quoted speech, which confirms the initial hypothesis.

		Mean Square	F	Sig.
Non-quoted speech	Sphericity Assumed	.255	2.716	.108
Quoted speech	Sphericity Assumed	4.966	18.902	.000
Questions	Sphericity Assumed	1.425	15.043	.000

Table 4: Tests of within-subjects effects

If we look at the results of the sphericity assumed test for quoted speech, things are different. The mean square is higher than for non-quoted speech (MS=4.966) and so is the F-ratio (F=18.902). The p-value does not exceed 0.05 (p=0.000), which means that there is a significant difference in MOS between the modified voice and the baseline. Thus, the implication of the descriptive statistics and the initial hypothesis that the baseline is perceived as significantly more natural can be confirmed.

The same applies for the voices when they are used in the domain of questions. The hypothesis was that some subjects might prefer the modified voice because its intonation rises towards the end of questions, and that others prefer the baseline because it sounds smoother, and that therefore no significant difference can be found. However, the descriptive statistics

showed that there is a preference for the baseline, and the sphericity assumed test proves that this difference is significant (MS=1.425, F=15.043, $p < 0.05$). Therefore, the initial hypothesis cannot be confirmed and further research is essential to achieve a good modelling of questions with rising intonation.

	Mean Square	F	Sig.
Experience (non-quoted speech)	.086	.119	.732
Experience (quoted speech)	1.088	1.374	.249
Experience (questions)	.533	.619	.436

Table 5: Tests of between-subjects effects

If we look at the effects of the factor experience on the rating of the voices, it is obvious that experience has no significant effect on the perception of naturalness of non-quoted speech (MS=0.086, F=0.119, $p > 0.05$), quoted speech (MS=1.088, F= 1.374, $p > 0.05$) and questions (MS=0.533, F=0.619, $p > 0.05$). However, the two groups of inexperienced and experienced listeners are difficult to compare since only 6 speech synthesis experts took part in the evaluation, which is not a representative sample of the population. It is therefore questionable if the results can be relied on, or if they would be different if more experienced listeners had taken part in the evaluation.

5.4 Evaluation of liveliness

5.4.1 Methodology

In the evaluation of liveliness, the subjects listened to two voices per page that said the same sentence. Then, the subjects had to decide which of the voices sounded livelier. The options were: *voice A sounds livelier*, *voice B sounds livelier* and *The voices sound equally lively to me*. For the statistical test, the results of each subject were transformed into scores and added up. If voice A sounded livelier, it received a score of 1 and the less lively voice a score of 0. If the two voices were perceived as equally lively, they both received a score of 1. As opposed to the MOS in the analysis of naturalness, the absolute scores were kept here for a better understanding. The scores of the subjects were compared with a one-way repeated measures ANOVA to analyse the within-subjects and between-subjects effects. Again, the tests were the same for non-quoted speech, quoted speech and questions. 20 recordings of each voice were played in each sub-section as well, but since two voices were presented on one page, each sub-section involved only 10 ratings.

What should be noted is that the interpretation of liveliness is very subjective and it might

differ between speakers. An exact definition of liveliness can therefore not be given, but what can be said is that a lively voice can be regarded as having a mixture of different properties. A lively voice should not be monotone, sound reasonably natural and be expressive. What is problematic is how the subjects see the concept of liveliness and how their concept influences the rating of the voices. A definition of what constitutes liveliness was not given to the subjects, but might have been helpful in their accomplishment of the task.

5.4.2 Results

The hypothesis was that when conveying non-quoted speech, the modified voice and the baseline do not differ significantly in terms of liveliness for the same reasons that were given in section 5.3.2. Because of the fact that only subtle differences between the voices are perceivable, it is difficult to say that one of them sounds livelier than the other. This is expected to be the case for both experienced and inexperienced listeners.

What quoted speech is concerned, the hypothesis was that the modified voice will win over the baseline voice. The higher pitched voice has a buoyant prosody and is quite expressive. The baseline voice is expressive as well, but it is not noticeable that characters are mimicked when the voice is applied in quoted speech. In contrast to the baseline, the higher pitch of the modified voice might be regarded as more vivacious. However, if subjects see naturalness as an inherent part of liveliness, their concepts of liveliness and naturalness might interfere and it is possible that they show a preference for the baseline. The same applies to questions. Although the rising intonation of the modified voice might be perceived as livelier, the fact that it has a buzzing quality and does not always sound smooth and stable, might cause the subjects to opt for the baseline voice. Nevertheless, the hypothesis was that for both quoted speech and questions, the modified voice is regarded as significantly livelier.

5.4.2.1 Descriptive statistics

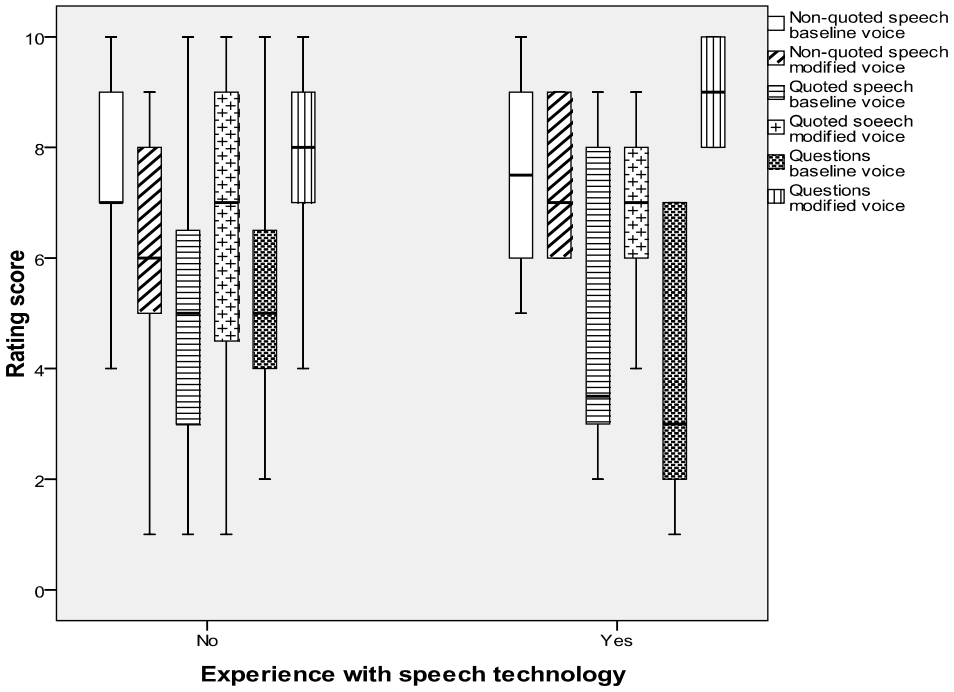

Fig. 15: Rating scores for liveliness

The descriptive statistics in Figure 15 show that there is a considerable difference in the perception of liveliness between the baseline voice and the modified voice for inexperienced listeners (the means are 7.53 and 6.13 respectively). This is surprising as there is indeed a small difference between the voices, but this would rather suggest the opposite outcome, namely that the modified voice should be regarded as livelier because of the richer prosody. The results of the experienced listeners emphasize that the difference between the voices is rather small, but they also seemed to regard the baseline voice as livelier (the means are 7.5 for the baseline voice and 7.3 for the modified voice). A close look at the individual data points showed that this was the case for the majority of participants. However, two experienced listeners rated the modified voice as livelier, with considerable differences between the scores of the baseline and the modified voice. This demonstrates that some experienced listeners perceived a difference in liveliness, but more participants would have been required to see which voice is preferred in a larger population.

The scores for liveliness differ considerably when it comes to quoted speech. It is noticeable that the lower bounds of the rating scores for the modified voice are much higher than for the baseline voice, especially for experienced listeners. What is furthermore obvious is that some experienced subjects seemed to regard the baseline as very lively as well, and a look at the the individual data points showed that two subjects considered the baseline as

livelier, while the remaining experienced subjects opted for the modified voice. The mean rating scores for the baseline are 5.03 and 4.83 and for the modified voice 6.5 and 6.83 for inexperienced and experienced listeners respectively.

With regard to questions, the differences in liveliness are even more substantial. The mean rating scores of the baseline voice are 5.16 and 3.83, while the mean rating scores of the modified voice are 7.78 and 9.0 for inexperienced and experienced listeners respectively. There is not only a noticeable difference between the scores within the two groups, but also between them: the inexperienced listeners rated the questions of the baseline voice as livelier than the experienced listeners, and for the modified voice it was the opposite case. A reason for this might be that the experienced listeners wanted to draw a sharper distinction between the baseline voice and the modified voice in terms of liveliness in questions, and that they also had a clearer concept of what it means for a voice to sound lively. The definition of liveliness was not quite clear to a large number of inexperienced listeners as they reported in a questionnaire at the end of the survey. This confusion is reflected by the rating scores which are close together, as opposed to the scores of the experienced listeners which are farer apart. Whether or not the differences reported here are significant will be answered in the following section.

5.4.2.2 Inferential statistics

Once more, a one-way repeated measures ANOVA was applied to investigate if the rating scores of the baseline voice and the modified voice differ within subjects. Furthermore, it was examined whether experience with speech technology has a significant effect on the judgement of liveliness. Again, Levene's test was not significant, so that the statistics in Table 6 can be trusted:

		Mean Square	F	Sig.
Non-quoted speech	Sphericity Assumed	6.250	1.688	.202
Quoted speech	Sphericity Assumed	30.397	2.637	.133
Questions	Sphericity Assumed	153.373	27.311	.000

Table 6: Tests of within-subjects effects

As it was expected from the descriptive statistics, the sphericity assumed test of within-subjects effects proved that the difference in rating scores between the baseline voice and the modified voice is not significant ($MS=6.25$, $F=1.688$, $p > 0.05$) in non-quoted speech. This confirms the initial hypothesis, in which it was claimed that the subtle differences in liveliness will not be perceived by the subjects. Moreover, although the descriptive statistics suggested

that there is a difference between the baseline voice and the modified voice in quoted speech, this difference is not significant either (MS=30.397, F=2.637, $p > 0.05$). Therefore, the initial hypothesis that the modified voice sounds significantly livelier in quoted speech than the baseline voice cannot be confirmed.

However, the voices differ significantly in liveliness what questions are concerned. The mean square and the F-ratio are substantially higher than in the previous within-subjects tests (MS=153.373, F=27.311) and the p-value is significant ($p < 0.05$). In this case, the initial hypothesis that the modified voice sounds significantly livelier than the baseline voice is supported. Table 7 shows whether the two groups of inexperienced and experienced listeners differed significantly in rating the voices:

	Mean Square	F	Sig.
Experience (non-quoted speech)	3.500	1.555	.220
Experience (quoted speech)	.46	.44	.834
Experience (questions)	.027	.016	.901

Table 7: Tests of between-subjects effects

If we look at the effects of the factor experience on the rating of the voices, it is again the case that experience has no significant effect on the perception of naturalness of non-quoted speech (MS=3.5, F=1.555, $p > 0.05$), quoted speech (MS=0.46, F= 0.44, $p > 0.05$) and questions (MS=0.027, F=0.016, $p > 0.05$). It can therefore be stated that there is a significant difference between the liveliness of the voices in questions, and that this is the case for both experienced and inexperienced subjects, as the descriptive statistics clearly show.

5.5 Evaluation of commercial viability

5.5.1 Methodology

In the evaluation of the commercial viability of the voices, the subjects listened to two voices on each page and had to decide which of the voices they would buy as their personal story-teller. The options were: *would buy none of them*, *would buy voice A*, *would buy voice B*, *would buy both*. For the statistical tests, these options were transformed into scores. If a subject would buy none of the voices, both voices received a score of 0. If a subject would buy voice A but not voice B, voice A received a score of 1 and voice B a score of 0 and vice versa. If a subject would buy both voices, both voices got a score of 1. The higher the score is for a voice, the more likely it is to be bought. Once more, the scores of the subjects were compared with a one-way repeated measures ANOVA to analyse the within-subjects and between-subjects effects. Again, the tests were the same for non-quoted speech, quoted speech and questions, and ten recordings were played for each voice.

5.5.2 Results

Whether or not a voice is likely to be bought is difficult to say, and this shall not be the main concern in proposing a reasonable hypothesis. It is more straightforward to say, which voice is more likely to be bought than the other, and this shall be the focus of the statistical analysis. Because of the fact that the subtle differences between the baseline voice and the modified voice are barely perceivable, it is hypothesised that the consumer behaviour will not be affected by these differences and that it is equally likely that the subjects buy the baseline or the modified voice.

With regard to quoted speech, opinions might be very diversified depending on whether the subjects prefer a more natural or a livelier voice, but the hypothesis is that the baseline voice will attract more potential buyers. First of all, naturalness is an essential feature a synthetic voice should have, and, as was already pointed out before, the modified voice needs to be improved what higher pitched speech is concerned.

The modified voice lacks further refinement in the domain of questions. While some questions sound pretty good, others have an unnatural intonation, which is due to an insufficient amount of training data. Therefore, the hypothesis is that the baseline voice is commercially more viable than the modified voice when it comes to questions.

5.5.2.1 Descriptive statistics

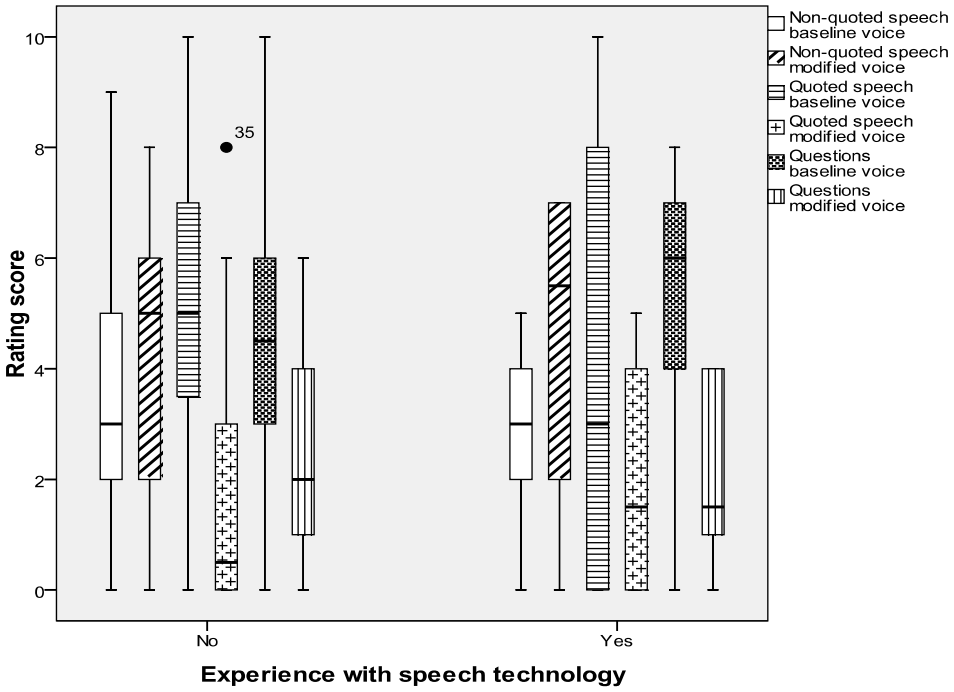

Fig. 16: Rating scores for commercial viability

Figure 16 shows that, in contrast to the previous descriptive statistics for naturalness and liveliness, the variety of scores that was given to the voices is much more varied. This is especially noticeable for the inexperienced subject group whose scores range from very low to very high for the baseline voice in all domains and the modified voice in non-quoted speech. The experienced subject group was overall more cautious and tended to give lower scores, i.e. on average, fewer experienced listeners than inexperienced listeners would buy the voices.

What is striking is that there is a considerable difference of commercial viability between the baseline and the modified voice in non-quoted speech for both inexperienced and experienced listeners. The mean rating scores for the baseline voice are 3.50 and 2.83 and for the modified voice 4.28 and 4.50 for inexperienced and experienced subjects respectively. There was not such a big difference between the rating scores of the baseline and the modified voice in the tests of naturalness and liveliness so that it was assumed that the subjects did not hear a major difference between both voices. However, here it can be seen that the subjects apparently perceived a difference nevertheless as the modified voice attracts more potential buyers. This suggests that this difference is not due to the naturalness or liveliness of the voice, but caused by another aspect that was not tested.

The commercial viability of quoted speech differs between the subject groups as well. Both groups agree that they would rather not buy the higher pitch aspect of the modified voice

(the mean rating scores are 1.59 for inexperienced and 2.0 for experienced subjects). The opinions seem to differ for the baseline voice because the range of scores that was given by the experienced subjects is quite large. Looking at the individual data points showed that this is the case because two subjects did not want to buy the baseline voice for quoted speech at all, and two other subjects were very attracted to buy it. The average scores for inexperienced and experienced subjects do, however, not differ considerably: the mean rating scores for the baseline are 4.88 for inexperienced listeners and 4.0 for experienced listeners.

What questions are concerned, the results of the two subject groups demonstrate that the baseline voice is once more commercially more viable. On average, the inexperienced listeners gave the baseline voice a score of 4.63, and the experienced listeners a score of 5.17. For the modified voice, the mean rating scores are lower, namely 2.38 and 2.0 for inexperienced and experienced listeners respectively. This confirms the assumption that the baseline voice is more attractive to buy when it comes to quoted speech and questions. However, the fairly low average score also show that the voices need to be improved if a large number of potential buyers shall be satisfied. If the differences in rating scores are significant will be shown in the analysis of the inferential statistics.

5.5.2.2 Inferential statistics

A further time, a one-way repeated measures ANOVA was applied to examine if the rating scores of the baseline voice and the modified voice differ within subjects. In addition to that it was analysed whether experience with speech technology has a significant effect on the buying behaviour of potential speech technology users. Again, Levene's test was not significant, and the statistics in Table 8 are feasible:

		Mean Square	F	Sig.
Non-quoted speech	Sphericity Assumed	15.138	3.356	.075
Quoted speech	Sphericity Assumed	70.463	9.785	.003
Questions	Sphericity Assumed	74.123	23.322	.000

Table 8: Tests of within-subjects effects

The results of the sphericity assumed within-subjects test for non-quoted speech show that the difference between the baseline and the modified voice is considerable, but not significant ($MS=15.138$, $F=3.356$, $p > 0.05$), which confirms the initial hypothesis. However, the significance value ($p=0.075$) is extremely near to the threshold and it is questionable that it happened by chance that the subjects were more attracted to buy the modified voice. Yet, more research is necessary to find out what aspect of the modified voice was more attractive

to the subjects.

Quoted speech is significantly more commercially viable when it is produced by the baseline voice (MS=70.463, F=9.785, $p < 0.05$). This is also the case for questions (MS=74.123, F=23.322, $p < 0.05$). These results support the initial hypotheses that the modified voice is less likely to be bought for the domains of quoted speech and questions, and that these domains need to be further improved.

	Mean Square	F	Sig.
Experience (non-quoted speech)	.507	.082	.776
Experience (quoted speech)	.555	.101	.753
Experience (questions)	.070	.012	.914

Table 9: Tests of between-subjects effects

If we look at the effects of the factor experience on the rating of the voices, it can be seen that experience has no significant effect on the commercial viability of non-quoted speech (MS=0.507, F=0.082, $p > 0.05$), quoted speech (MS=0.555, F= 0.101, $p > 0.05$) and questions (MS=0.070, F=0.012 $p > 0.05$). Therefore, it can be pointed out that there is a significant difference between the commercial viability of the two voices in quoted speech and questions, and that this is the case for both experienced and inexperienced subjects.

5.6 Summary and discussion of results

The preceding evaluation has shown that the baseline voice and the modified voice are significantly different when it comes to naturalness in quoted speech and questions. Furthermore, the liveliness of both voices differs significantly in questions. Finally, it was demonstrated that the commercial viability of the two voices is substantially different for non-quoted speech and significantly different for quoted speech and questions.

Although it was intended to build a synthetic voice that performs well in character mimicking and asking questions, this was not accomplished because the baseline voice was preferred in both domains. This might have been different if the modified voice had not been that buzzy and less creaky when conveying higher pitched speech. Further research is necessary to achieve a better modelling of the f0 ranges. This might be accomplished by using more training data of higher pitched speech or by trying another acoustic modelling technique. The same applies to questions. In contrast to declarative sentences, the training data included only few questions. However, finding an audio book with a sufficient amount of questions is difficult. A solution might be to take questions from various male speakers and

apply a speaker adaptation technique such as SAT (cf. Gibson and Byrne 2011).

A further problem was that the inexperienced listeners had problems rating the voices. Each subject had to answer a questionnaire at the end of the listening test where they could note down their suggestions and difficulties. About a third of the inexperienced subjects regarded the rating of naturalness as problematic because they were either not exactly sure what the characteristics of natural speech are, or they considered the scores from 1-5 as hard to differentiate. The rating of liveliness was even more complicated for the inexperienced subjects. Nearly half of the inexperienced listeners stated that they did not know what the concept of lively speech should include. Furthermore, they remarked that some of the livelier sentences, presumably from the modified voice, sounded less natural and that this interfered with their liveliness judgement. Therefore, it would have been helpful to define the concept of liveliness before the listening test.

However, most of the experienced listeners did not have any problems with the liveliness judgement and none of them found it difficult to rate the naturalness of the voices. This suggests that the experimental design needs to be improved in terms of making the evaluation easier for the target group of synthetic voices, which are usually people who are not experienced with speech technology.

6 Conclusion

At the beginning of this paper, the theory of HMM-based speech synthesis was introduced to the reader to convey an impression of how voice building was conducted with the speech synthesis system HTS. After the training data was described and analysed in detail in chapter 4, the modifications that were built into the speech synthesis system were outlined and it was argued why these modifications are regarded as necessary to improve the synthetic voice.

However, the evaluation in chapter 5 showed that these modifications are not accepted by the audience and that the baseline voice wins over the modified voice in the majority of domains. There is a slight indication that the subjects regarded the modified voice as better than the baseline in non-quoted speech, but not in terms of naturalness or liveliness. What causes this preference is subject to further research. It might be the case that character mimicking in a higher pitched voice as well as questions with a rising intonation are accepted by the listeners when these domains sound as smooth and stable as non-quoted speech. Once more, further research is necessary to accomplish this.

7 References

Arifianto, D. and T. Kobayashi. “Voiced/unvoiced determination of speech signal in noisy environment using harmonicity measure based on instantaneous frequency”. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, vol. 1, pp. 877-880, Philadelphia, Pa, USA, March 2005.

Bennett, C.L. and Alan Black. “Prediction of Pronunciation Variations for Speech Synthesis: A Data-driven Approach”. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, Pennsylvania, 2005.

Benoît, Christian, Martin Grice and Valérie Hazan. “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences”. *Speech Communication*, vol. 18, pp. 381-392. 1996

Braunschweiler, N., Gales, M.J.F., Buchholz, S. "Lightly supervised recognition for automatic alignment of large coherent speech recordings", in Proc. of Interspeech, Makuhari, Chiba, Japan, pp.2222-2225, 2010.

Chomphan, S. “Analysis of Decision Trees in Context Clustering of Hidden Markov Model Based Thai Speech Synthesis”. *Journal of Computer Science*, vol. 7, pp. 359-365, 2011.

Field, Andy. *Discovering Statistics using SPSS*. London: Sage Publications. 2005.

Gibson, M. and W. Byrne. “Unsupervised Intralingual and Cross-Lingual Speaker Adaptation for HMM-Based Speech Synthesis Using Two-Pass Decision Tree Construction”.*IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 895-904, 2011.

Jurafsky, Daniel and James H. Martin. *Speech and Language Processing*. Pearson Education: Upper Saddle River, New Jersey. 2009.

Kawahara, H, H. Katayose, A. Cheveigné, and R. Patterson. “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity”. *Proc. EUROSPEECH 1999*, pp. 2781–2784, Sep. 1999.

King, Simon. *A beginners' guide to statistical parametric speech synthesis*. University of Edinburgh: The Centre for Speech Technology Research. 2010.

Murphy, Kevin P. *Hidden semi-Markov models*. Technical report, MIT AI Lab, 2002.

Prahallad, Kishore. *Automatic Building of Synthetic Voices from Audio Books*. School of Computer Science: Carnegie Mellon University, Pittsburgh. 2010.

Renals, Steve. *Lecture Automatic Speech Recognition*. School of Informatics: University of Edinburgh, 2011.

D. Talkin. “A robust algorithm for pitch tracking (RAPT)”. Speech Coding and Synthesis. W.

Kleijn and K. Paliwal, Eds. Elsevier, pp. 495–518, 1995.

Tokuda, K, H. Zen and A.W. Black. “An HMM-based speech synthesis system applied to English”. *Proc. of 2002 IEEE SSW*, Sept. 2002.

J. Yamagishi, Z. Ling, and S. King.“Robustness of HMM-based Speech Synthesis”. *Proc. of InterSpeech*, pp.581—584, 2008.

Yamagishi, Y, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King and S. Renals. “A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis”. *IEEE Audio, Speech, & Language Processing,* vol.17, no.6, pp.1208-1230, August 2009.

Zen, Heiga, Keiichi Tokuda and Tadashi Kitamura. “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences”. *Computer Speech & Language*, vol.21, no.1, pp.153-173, Jan. 2007(a).

Zen, Heiga, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan Black and Keiichi Tokuda. “The HMM-based speech synthesis system (HTS) version 2.0”. *Proc. 6th ISCA Workshop on Speech Synthesis (SSW-6)*, August 2007(b).

Zen, Heiga, Keiichi Tokuda, and Alan W. Black. “Statistical parametric speech synthesis”. *Speech Communication*, vol.51, no.11, pp. 1039-1064, November 2009.

Zen, Heiga and Mark J. F. Gales. “Decision tree-based context clustering based on cross validation and hierarchical priors”. *Proc. ICASSP*, pp.4560-4563, Prague, Czech, May 2011.