8th ISCA Speech Synthesis Workshop • August 31 – September 2, 2013 • Barcelona, Spain

Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS

Norbert Braunschweiler, Langzhou Chen

Toshiba Research Europe Ltd., Speech Technology Group, Cambridge, United Kingdom
{norbert.braunschweiler,langzhou.chen}@crl.toshiba.co.uk

Abstract

The presence of inhalation breaths in speech pauses has recently attracted more attention especially since the focus of speech synthesis research has shifted to prosodic aspects beyond a single sentence, as, for instance in the synthesis of audiobooks. Inhalation breath pauses are usually not an issue in traditional speech synthesis corpora because they typically use single sentences of limited length and therefore pauses including inhalation breaths rarely occur or they are deliberately avoided during recording. However, in readings of large coherent texts like audiobooks, there are often inhalation breaths, particularly in publicly available audiobooks. These inhalation breaths are relevant for the modelling of pauses in audiobook synthesis and can cause a reduction in naturalness when un-modelled. Therefore this paper presents a method to automatically classify pauses into one of four classes (silent pause, inhalation breath pause, noisy pause, no pause) for improved pause modelling in HMM-TTS.

Index Terms: inhalation breaths, pauses, speech synthesis, HMM-TTS, classification

1. Introduction

Inhalation breaths in speech pauses have not attracted much attention in the history of speech synthesis research. Exceptions are, for instance [1] who showed that the inclusion of inhalation breaths before an utterance improved the recall of synthesized sentences and [2] who indicated that their inclusion in a limited domain synthesizer improved its naturalness.

One of the reasons for this lack of interest is certainly its relatively small impact on quality and naturalness. Another reason is undoubtedly the limitations of traditional TTS training material which is typically limited to single sentences recorded in controlled conditions. In such a recording style the presence of inhalation breaths was typically avoided to achieve a homogenous representation of speech pauses as silent pauses. However, with the move to study prosodic aspects beyond the single sentence and in particular in the domain of audiobook synthesis inhalation breath noises are becoming more important.

One issue is the presence of pauses including inhalation breath noises in the training data. If un-modelled they can degrade synthesis quality, because often silent pauses and pauses including inhalation breath noises or other articulatory noises are considered as a single unit.

The reason for the presence of inhalation breaths in the training material comes with the move to use speech corpora which include prosodic features beyond a single sentence. Audiobooks are ideal for this because they are typically based on a coherent text and include large amounts of speech from a single or multiple speaker(s). However, unless inhalation breaths and other articulatory noises within pauses are deliberately removed, audiobooks include them. This has been observed particularly in the domain of publicly available audiobooks, as used, for instance in the Blizzard Challenge 2012 [3], but also recently studied on a French audiobook [4].

To use publicly available audiobooks as training material for speech synthesis purposes the speech first needs to be segmented and aligned with its corresponding text. This task was provided for the Blizzard Challenge 2012 by the lightly supervised sentence alignment and selection approach [5]. This approach automatically selects individual sentences from an audiobook for which it detects a match between recognition and expected text. The output of the lightly supervised approach is a set of sentences for which there is an audio file and a corresponding text file. Sentence sized audio files are cut out from the larger, typically chapter sized audio files. Cutting points are the mid points of any pause between sentences or the end of the sentence final phone if no pause is present. As such, the sentence sized audio files often include parts of inhalation breaths, articulatory noises (lip smacks, clicks, etc.) or other background noise in their leading/trailing parts as well as in their sentence internal pauses.

Many current synthesis frameworks are still very much limited to work on single sentences only. When the intervals of audio labelled as leading/trailing silence or sentence internal pause are not as clean as in the (controlled) studio recordings it can adversely affect speech synthesis quality. The synthesised leading and/or trailing silences can include inhalation breaths or other noises or a mixture of these and the same can happen in sentence internal pauses.

To avoid this problem and to provide the basis for modelling inhalation breath pauses in synthesis, an approach is presented which classifies any automatically labelled pause into one of four categories (silent pause = *pau*, inhalation breath pause = *paub*, noisy pause = *paun*, no pause = *no_pau*). The classification result can then be used during training and synthesis.

This paper is structured as follows: first the results of a pilot study are presented which used natural speech to test the impact of inhalation breath pauses on perceived naturalness. Then a new approach to automatically classify pauses into four subclasses is presented and evaluated. This is followed by the application of the classification approach to a publicly available audiobook. Based on the classification results an HMM-TTS model is trained and compared with a system using a single pause model. The the discussion addresses related issues and possible future research directions. Finally the conclusion summarizes the crucial findings.

2. Pilot study: Impact of breath pauses on naturalness

To test the influence of inhalation breath pauses on perceived naturalness a pilot study was conducted using natural speech from a German TTS speech corpus. This corpus, read by a female voice talent, had a number of longer sentences as well as sentences originally recorded as part of paragraphs and therefore included some inhalation breaths.

A standard preference listening test was conducted contrasting presence vs. absence of inhalation breath pauses. The test was conducted via crowd sourcing using the CrowdFlower website and subjects in Germany. The analysis included automatic cheat detection [6] and standard paired t-test for statistical significance calculation. The test used 40 sentences each of them including at least one inhalation breath pause. These evaluation sentences were selected from mixed text genres like news, navigation, audiobook, etc. and were relatively long sentences. The average number of words in the evaluation sentences was 30.3 ± 14.0. There was a total of 174 pauses (79% *paub*, 20% *pau*, 1% *paun*) and the average number of pauses was 4.3 ± 2.4 with a mean pause duration of 354.9 ± 155.3 ms.

For the preference test all inhalation breath pauses and noisy pauses as well as any noisy leading/trailing silences were manually "silenced" by using wavesurfers (http://www.speech.kth.se/wavesurfer/) "Amplify" function to reduce the amplitude of any (inhalation breath) noise to almost zero, i.e. to make it in-audible and effectively create a silent pause while still not silencing it completely. Complete "silencing" was avoided because it can result in un-natural transitions from speech to pause and vice versa. The amplitude flattening was carefully applied to avoid touching any transition to and from the pause as long as the (inhalation breath) noise could be made inaudible. The effect of the amplitude flattening was checked auditorily and repeated application was conducted when there was still audible (inhalation breath) noise.

The test compared the natural sentences (henceforth "Natural") against the sentences with amplitude flattened pauses (henceforth "No_breath"). Subjects were asked: "Indicate which of two speech sound files sounds more natural?". 660 pair stimuli were evaluated by 30 subjects. The results are presented in Table 1 and show no clear preference but a very small and statistically non-significant tendency to prefer the natural stimuli.

Table 1: *Result of preference listening test in pilot study.*

Natural	No_breath	none	p-score
35.9%	31.8%	32.3%	0.147

This result indicates that there is some impact of inhalation breath pauses on perceived naturalness, but the impact is small and statistically non-significant.

Coming back to inhalation breath pauses in speech synthesis, the first issue to address is the detection of inhalation breath pauses in the training data. The next section presents an approach to automatically classify pauses into the mentioned four sub-classes (*pau*, *paub*, *paun*, *no_pau*) based on pause labels automatically annotated by forced alignment.

3. Automatic classification of pauses

Figure 1 depicts the basic steps from a speech corpus to an acoustic model for HMM-TTS and shows the location of the pause classification method introduced below.

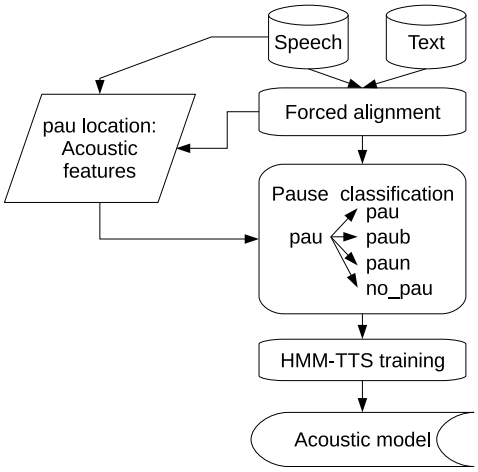

Figure 1: *Flow diagram from speech and text to acoustic model via automatic pause classification.*

Typically a TTS corpus consists of speech and corresponding text files. These are input to the forced alignment module which provides alignments of any phones and pauses. The locations of these pauses are used by the pause classifier to extract acoustic features from the corresponding stretches of speech. The output of the classifier is then used in HMM-TTS model training to train an acoustic model. The next section will give a brief explanation of the classification method.

3.1. Classification method

To enable the separation of silent pauses and pauses filled with inhalation breaths or other noise a method was developed which classifies each speech pause into one of the following four classes:

- *pau* - silent pause
- *paub* - pause including inhalation breath
- *paun* - pause including any other noise than paub
- *no_pau* - no pause

The *no_pau* class was introduced to account for the fact that some of the automatically annotated pauses are incorrect pause insertions and need to be deleted from the alignments.

The same classification is also provided for any leading and trailing silence in a given speech file with the exception that no silence will be deleted but a warning message will be printed when the classifier identifies a silence which is deemed to include speech.

The classification method uses a set of acoustic features and the type of the preceding and following phone to perform the classification. The features are scored by a set of rules and then threshold values are used for the final classification.

Feature values are extracted on a frame level and various statistics are calculated for either the whole pause or specified parts of it. The set of acoustic features is described in section 3.3. The next section presents some considerations about the input data to the classifier which are the automatically aligned phones and pauses.

3.2. Automatic alignment of pauses

Unless there is manually annotated data available, automatic phone and pause alignments are usually the basis for an HMM-TTS voice. Typically pauses are labelled automatically as part of forced phone alignments using HMM models. These models can be trained on a given corpus by a flat start approach or using speaker independent models trained on large, multi-speaker corpora. In the presented approach the automatic pause alignments were created by a flat start method which - after some iterations of maximum likelihood training - introduces a one state short pause tee-model, which is tied to the centre state of the silence model. Models are then iteratively refined and the short pause model is changed to a three state tee-model, where each state is tied to the corresponding state in the silence model.

Since this short pause model shares states with the silence model the presence of silences including inhalation breaths affects the performance of the short pause model. One issue with automatic pause alignments is the precision of the pause boundaries. Typical problems are the inclusion of parts of neighbouring sounds into the pause itself. This issue will be addressed in section 3.3 in more detail. In the presented approach no attempt was made to alter the HMM model topology. The automatic silence and pause alignments were used as provided by this method. The next section presents the acoustic features used in the classifier in more detail.

3.3. Acoustic features

At first glance the separation of silent pauses and inhalation breath pauses seems to be a straightforward approach. Silent pauses are expected to contain no voicing, very little and constantly low energy, whereas pauses filled with inhalation breaths are expected to include no voicing as well, but a higher energy level and a distinct spectral energy distribution.

To see whether these expectations can be used to classify pauses a small subset of pauses, including all four classes, were analyzed with respect to their acoustic features including f0 (for voicing detection), RMS-amplitude, and spectral energy distribution.

During this analysis it was noticed that the pause boundaries placed by the automatic aligner were not always precise but often included parts of neighbouring phones. When the preceding phone was a stop, for instance, the stop releases could be partly or completely subsumed within the pause. Depending on the type of phone before the pause the onset part of the pause could include voiced or unvoiced parts having various non-silence spectral intensities. Sometimes more than half the pause was covered by the preceding sound. Similar observations were made for the pause ends. This meant that any method using acoustic features to classify pauses had to account for these imperfectly placed pause boundaries.

Figure 2 shows schematic representations of the amplitude tracks for the three pause classes *pau*, *paub*, *paun*, indicating onset and offset parts, which can include features from neighbouring phones and which can vary according to the precision of automatically placed boundaries.

Given the analysis a set of acoustic features was chosen based on the considerations to detect presence/absence of voicing, levels of energy and distribution of spectral energy within pauses - to enable a reliable classification of each pause into one of the four classes mentioned.

The ESPS-tools *get_f0* and *sgram* were used to calculate f0, RMS-amplitude, and a series of FFT's providing information about spectral energy distribution. The *get_f0* tool was used

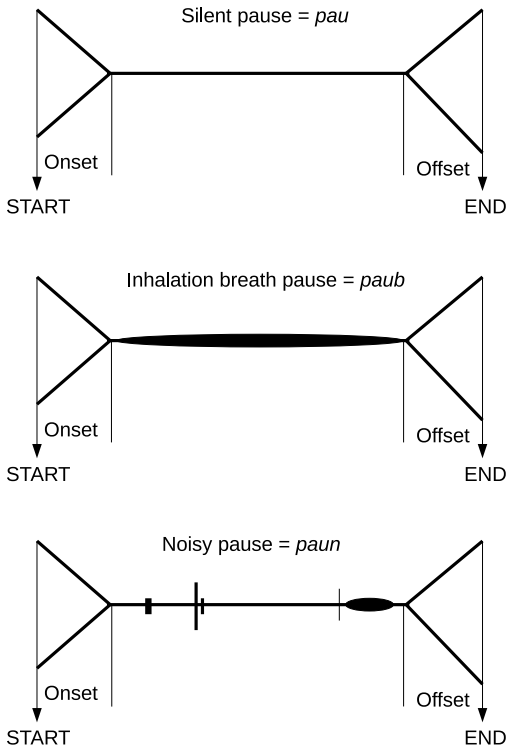

Figure 2: *Schematic representation of amplitude tracks of pause classes: pau, paub, and paun.*

with standard settings, i.e. a frame step of 10 ms. The *sgram* tool was called with the wideband spectrogram option (-m wb) which uses a frame step of 2 ms [7].

The automatically labeled start and end timestamps of each pause were used to extract these acoustic features from the corresponding stretches of each sentence sized audio file. Using automatically placed timestamps is important because the classifier is expected to work for exactly these timestamps which are not always precise as mentioned before. In case the pause duration was less than 30 ms no f0 and RMS-amplitude was calculated because the RMS value of each record is calculated on a 30 ms hanning window [7] and feature values for these were set to zero. Table 2 lists the acoustic features.

Features were introduced which observed particular parts of the pause, e.g. the first/last 20% or the first/second half. Previous internal experiments revealed that HMM-TTS synthesis quality is relatively insensitive against imprecise phone and pause boundaries which meant that imprecise pause boundaries are to some extend tolerated and should not trigger an incorrect classification or an incorrect deletion of a pause.

The features chosen include features which calculate average values across the whole pause as well as average values across defined parts of the pause. Additionally the spectral energy distribution across the full frequency range was also calculated for particular bands (low, lowmid, mid, high) based on the findings for inhalation breath pauses in the training data. Threshold values for low spectral power (perc_spec_pwr_low) and non-low spectral power (perc_spec_pwr_high) were defined, also based on observations in the training material. For all three basic features (f0, RMS, spectral energy) its mean, maximum and minimum values were calculated across the whole pause to

get an idea about levels and variances of them.

In addition to the above mentioned observations it was also noticed that a few pauses were completely incorrect, e.g. labeled within speech without any pause nearby and a few were debatable whether to be labelled as pause or being left as part of neighbouring phones especially in the case of glottal stops following or in stop - stop sequences.

Table 2: *List of acoustic features.*

Feature name	Definition
pau_dur	Duration of pause in ms
perc_voiced	% of voiced frames
f0_mean	mean f0
f0_max	max f0
f0_min	min f0
RMS_mean	mean RMS
RMS_max	max RMS
RMS_min	min RMS
RMS_std	std. dev. of RMS
spec_pwr_mean	mean spectral power
spec_pwr_std	std. dev. of spec. power
perc_spec_pwr_low	% power $<$10
perc_spec_pwr_high	% power $>$100
perc_spec_pwr_high_onset	% power $>$100 in 1st 20%
perc_spec_pwr_high_offset	% power $>$100 in last 20%
perc_spec_pwr_high_1sthalf	% power $>$100 in 1st half
perc_spec_pwr_high_2ndhalf	% power $>$100 in 2nd half
spec_bands_mean	mean all spectral bands
spec_bands_mean_std	std. dev. all spectral bands
spec_bands_max	max of all spectral bands
spec_bands_min	min of all spectral bands
spec_bands_mean_low	mean [0, 900] Hz
spec_bands_mean_lowmid	mean [1600 - 2000] Hz
spec_bands_mean_mid	mean [5000 - 6000] Hz
spec_bands_mean_high	mean [6000 - 11025] Hz

3.4. Evaluation

The next section describes the training corpus used to select the acoustic features and to define the rules in the classifier.

3.4.1. The training corpus

For training the automatic pause classifier 2291 sentences from a German speech synthesis corpus (the same as in the pilot study) were used including 4563 hand labelled pauses. This data was split into 90% training and 10% test sets. The hand labelling included the inspection of each sentence and each included pause (listening and visualising its waveform), its relabelling as either *pau*, *paub*, *paun* or its deletion. When the labeller noticed a missing pause it was inserted. Also timestamps of pauses were corrected when necessary.

The original automatic markup (henceforth called AUTO_ORIG) consisted of 2354 sentences including 4512 pauses. Because the human labeller deleted and added pauses there were 108 sentences removed in the hand markup (henceforth called HAND).

Table 3 shows how the HAND labelled pauses compare with the AUTO_ORIG pauses - for the 2354 sentences originally including a pause in the automatic markup. 75.4% are silent pauses, 15.6% include inhalation breaths, 3.9% are pauses including any other audible noise than an inhalation breath and 5% were deleted. The last row in Table 3 shows the number of pauses added by the human labeller.

Table 3: *Comparing pauses in AUTO_ORIG and HAND.*

	AUTO_ORIG	HAND
# pauses	4512	4509
pau	100%	75.4%
paub	-	15.6%
paun	-	3.9%
deleted	-	5.0%
added	-	224

The mean durations and standard deviations of the three pause classes are shown in Table 4. As can be seen the classes show distinctive mean durations: inhalation breath pauses are longest, followed by silent pauses, which have the highest standard deviation, and the noisy pauses being the shortest on average.

Table 4: *Mean pause durations and standard deviations in the training corpus.*

Pause	Mean [ms]	StdDev [ms]
pau	210.7	134.6
paub	393.1	104.2
paun	139.4	120.8

3.4.2. Results

To measure the accuracy of the pause classifier 10-fold cross-validation was conducted. For each test set it was ensured that the four pause classes were proportionally represented as in the HAND data.

Average precision was 0.87 ± 0.014 and average recall 0.86 ± 0.014. This shows that the majority of pauses were correctly classified, while on average 5.9% of *pau* labels were falsely classified as *paub*, but only 1.1% of the *paub* labels were missed. This is probably caused to a large extent by the limitations to filter out incorrect pause boundaries. When neighbouring phones are part of the pause then the acoustic features become "polluted" by these and can result in incorrect classifications. Confusions of other classes are relatively small. While the deletions are similar in numbers to the deletions in HAND, 64.3% of the pauses deleted in HAND were correctly deleted the remainder was not deleted and 18.9% were falsely deleted. This can certainly also be mostly explained by the imperfect pause boundaries which could trigger a pause deletion when the acoustic features are strongly dominated by the speech parts and not by pause parts. The classification of a pause as *no_pau* was often an indication that the automatic alignment was either completely incorrect or largely incorrect, i.e. could be used to spot mis-alignments.

To conduct another test of the classifier on unseen data it was applied to a publicly available American English audiobook. Classified pause labels were then used in the training of an HMM-TTS model and compared with the standard, single pause system. This is described in the following section.

4. Synthesis with multi-pause labels

4.1. Classifying pauses in "A Tramp Abroad"

The classifier was applied to the publicly available audiobook "A Tramp Abroad" (librivox.org) written by Mark Twain and read by John Greenman. This audiobook was part of the training material used during Blizzard Challenge 2012 (http://festvox.org/blizzard/blizzard2012.html).

For this paper the subset of sentences selected at a 100% confidence interval (by the lightly supervised sentence selection and alignment tool [5]) was used consisting of 5052 sentences and including 8624 pauses in total. These pauses were all automatically aligned by the Toshiba internal automatic phone alignment tool. From the 5052 sentences 64.7% did contain at least one pause and 36.2% did not include a pause. Without pause classification the average pause duration was 299.9 ± 155.0 ms.

After classifying each pause into the four classes of *pau*, *paub*, *paun* and *no_pau* there are 8073 remaining pauses. That means, 6.4% were deleted and the remaining ones were classified as follows: *pau*: 15.8%; *paub*: 72.6%; *paun*: 5.1%.

As presented in Table 5, the average pause durations of the three pause classes showed a similar duration pattern as in the German TTS training corpus (*paub* >*pau* >*paun*). However, the mean duration of inhalation breath pauses was shorter than in the German speaker, indicating that there are more shorter inhalation breath pauses in "A Tramp Abroad", an observation which is in line with listening impressions on several samples.

However, there are many more inhalation breath pauses in "A Tramp Abroad" than in the German TTS training corpus. This is not surprising when listening to this data which shows that the reader of "A Tramp Abroad" inhaled quite frequently, whereas the voice talent of the German speech database tried to avoid inhaling during the single sentence prompts.

Table 5: *Mean pause durations and standard deviation in "A Tramp Abroad".*

Pause	Mean [ms]	StdDev [ms]	% of original
pau	218.3	136.7	15.8%
paub	352.7	126.5	72.6%
paun	146.4	94.5	5.1%

Because there are no hand annotated pause labels for "A Tramp Abroad" it was not possible to quantify the accuracy of the classifier on this corpus. However, visual and auditory inspection of some sentences showed that the classification worked reasonably well. The next section will test the impact of multi-pause labels on synthesis quality.

4.2. HMM-TTS training

To test the impact of the classified pauses in synthesis two listening tests were conducted. An HMM-TTS voice was trained on the audiobook "A Tramp Abroad" read by John Greenman, i.e. the same audiobook as used in section 4.1 for the automatic classification of pauses.

The training corpus contained about 9 hours of speech in 4.8K utterances (the remaining sentences were left out as test set) and was sampled at 16k Hz. The acoustic feature vectors included 40 mel-cepstral coefficients, logF0, 21 band aperiodicity together with their delta and delta-delta information. They were modelled by multi-stream, 5 state, left-to-right, multi-space probability distribution hidden semi-Markov models (MSD-HSMM). The full-context HMM states were generated by introducing the phonetic, segmental, prosodic and linguistic context information. Decision tree based state clustering was used for tying the full-context HMM states based on the minimum description length (MDL) principle.

Two set of MSD-HSMMs were trained based on two phone lists. One used a single label for silence and a single label for pause, the other used multiple labels for silences and multiple labels for pauses. In the second case, the silence was divided into 3 separate "phonemes", pure silence, inhalation breath silence and noisy silence, in the same way for the short pause. In the training process, each of them was modelled individually. They were considered as different context information when full-context HMM states were generated, and the questions about the type of non-speech events were also used in the process of decision tree growing. This way, not only the different types of non-speech events were explicitly modelled, but also their influence on neighbouring phonemes was explicitly considered in decision tree generation. Thus a better distribution sharing over full-context HMM states can be achieved.

4.3. Results of listening tests

The first preference test was designed to address the question whether the finer split of pauses affects synthesis quality. 25 sentences from the test set of "A Tramp Abroad" were selected, each of them including at least one pause. To avoid the impact of automatic pause prediction, the pauses which were automatically annotated by the forced alignment were used in case of the SINGLE_PAU system and the sub-classified version of them in the MULTI_PAU system. There were 45 pauses in the automatic alignments of these sentences. These 45 pauses were classified as follows: *pau*: 4, *paub*: 38, *paun*: 2, *no_pau*: 1.

Each sentence was synthesized by two different synthesizers: synthesizer SINGLE_PAU was trained using a single pause model and synthesizer MULTI_PAU used the sub-classified pause labels for training (as described in section 4).

Subjects were asked to indicate which of two speech sound files sounds better. The test was conducted with the crowd sourcing platform CrowdFlower using subjects in the USA. 512 pair stimuli (after discarding cheats) were rated by 42 subjects. Results are presented in Table 6.

Table 6: *Results of preference listening test comparing MULTI_PAU vs. SINGLE_PAU.*

MULTI_PAU	SINGLE_PAU	none	p-score
42.6%	49.4%	8.0%	0.061

There was a statistically non-significant difference between the two systems, but a small preference for the SINGLE_PAU system. This shows that there is no significant negative impact of the multi-pause classification, but also no improvement in quality.

The second listening test was designed to address the question whether the multi-pauses add to the naturalness perception. This time paragraphs were chosen to test the impact on longer stretches of speech including multiple pauses. 20 paragraphs were selected from the test set of "A Tramp Abroad" and pause positions were taken from the original reading and not predicted in order to avoid the impact of incorrect pause prediction. The 20 paragraphs included 60 sentences which in turn included 63

pauses which were classified as follows: *paub*: 44, *paun*: 3, *pau*: 8, *no_pau*: 8.

All 20 paragraphs were synthesized by the same HMM-models trained on the multi-pause classification. However, while system MULTI_LAB used the multi-pause labels also in synthesis, system SINGLE_LAB only used the silent pauses, therefore effectively representing a synthesizer only generating silent pauses.

By listening to each stimuli it was confirmed that the SINGLE_LAB stimuli included silent pauses, an indication that the classification worked well.

Subjects were asked to indicate which of two speech sound files sounds more natural. Again the test was conducted via crowd sourcing. 328 pair stimuli (after discarding cheats) were rated by 32 subjects. The results are shown in Table 7.

Table 7: *Results of preference listening test comparing MULTI_LAB vs. SINGLE_LAB.*

MULTI_LAB	SINGLE_LAB	none	p-score
58.5%	32.0%	9.5%	$<$0.001

Subjects significantly preferred system MULTI_LAB, showing the impact of multiple pauses in training and synthesis as opposed to a single, silent pause only synthesizer.

5. Discussion

The pilot study was comparing single sentences, although relatively long ones, however, it might be necessary to use paragraphs or even longer stretches of coherent speech, because then the presence of inhalation breath pauses might be more important and appear more natural to the listener. Inhalation breaths can add to the naturalness of speech particularly in an audiobook scenario. Here, they could help to realise a more expressive reading style, when modelled correctly, for instance when particular scenes occur, where inhalation breaths help to make the story more lively, and - if omitted - might result in a lack of naturalness.

The results of the listening tests with synthetic speech showed that a more fine grained pause classification can help to improve the naturalness, especially in longer texts. However, since the test was using pause locations originally placed by the human speaker the next step that is needed, is the prediction of the more fine-grained pause classes from text. A task that is known to be difficult in the speech synthesis world, see chapter 6.2 and 6.7 in [8] about phrasing and phrasing prediction respectively.

Another aspect is the capability of the system to synthesize natural sounding inhalation breaths. While the inhalation breaths synthesized with the Toshiba HMM-TTS did sound acceptable there is certainly potential for improvement.

Furthermore this work may provide the basis for subsequent work about the patterns of pause positioning and pause duration timing in coherent texts. By extending this work to account for the amount of detail observed in natural speech including inhalation breaths and possibly other pauses as well (i.e. “filled pauses” including laughter, vocatives, etc.) it might be possible to produce more natural sounding synthesis especially when synthesising large, coherent texts as in the audiobook scenario.

A possible extension of the current classifier would be to add the functionality to adjust pause boundaries. This could be beneficial in scenarios where pause boundary precision is more important, e.g. for unit selection systems.

6. Conclusions

This study investigated inhalation breath pauses and their influence on perceived naturalness in natural as well as in synthetic speech. A pilot study using natural speech, showed a small, but statistically non-significant preference for natural speech including inhalation breaths in pauses against a version which had the inhalation breaths silenced.

Following this study, a pause classifier was developed using a set of acoustic and phonetic features to classify each automatically labelled pause into one of four classes: silent pause, inhalation breath pause, noisy pause and no pause.

The approach was trained and evaluated on a German speech synthesis corpus and showed a good accuracy, especially with respect to the detection of inhalation breath pauses. The classifier was then applied to a publicly available American English audiobook and the classification results were used to train an HMM-TTS system which was compared against an HMM-TTS system trained on the same data, but only using a single pause model. Two listening tests were conducted, the first one testing the impact of the multi-pause labels on synthesis quality. No significant difference was found between systems. The second preference test addressed the question whether the multi pause labels improve the naturalness of synthesis. This time full paragraphs were evaluated and the multi-pause label system was significantly preferred against the single pause label.

The effect of having just silent pauses in speech synthesis can be more severe when synthesizing audiobooks and can add to the perceived naturalness of synthetic speech. Adequate modelling of pauses - either silent, with inhalation breaths or any other form - is important for non-monotonous and prosodically well-structured speech synthesis.

7. Acknowledgements

The authors would like to thank the teams behind Librivox.org and Gutenberg.org for hosting voluntarily produced audiobooks and out of copyright books. We would also like to express our gratefulness to John Greenman, who made his narration of “A Tramp Abroad” publicly accessible.

8. References

[1] D. H. Whalen, C. E. Hoequist, and S. M. Sheffert, “The effects of breath sounds on the perception of synthetic speech,” in *J. Acoust. Soc. Am. Volume 97, Issue 5*, 1995, pp. 3147–3153.

[2] S. Sundaram and S. Narayanan, “Spoken language synthesis: experiments in synthesis of spontaneous monologues,” in *Proc. of IEEE Workshop on Speech Synthesis*, 2002, pp. 203–206.

[3] S. King and V. Karaiskos, “The Blizzard Challenge 2012,” in *Proc. of Blizzard Challenge 2012 Workshop*, 2012.

[4] G. Bailly and C. Gouvernayre, “Pauses and respiratory markers of the structure of book reading,” in *Proc. of Interspeech*, 2012.

[5] N. Braunschweiler, M. Gales, and S. Buchholz, “Lightly supervised recognition for automatic alignment of large coherent speech recordings,” in *Proc. of Interspeech*, 2010, pp. 2222–2225.

[6] S. Buchholz, J. Latorre, and K. Yanagisawa, “Crowd sourced assessment of speech synthesis,” in *Crowd Sourcing for Speech Processing Applications to Data Collection, Transcription and Assessment*. Wiley & Sons, 2012.

[7] ESPS/waves+, “Manuals of product release 5.3,” in *Entropic Inc., Washington, DC*, 2001.

[8] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.