Pitch-Effects in Diphone Recording: Are Logatomes inappropriate?

Ulrich Reubold, Alexander Steffen

Department of Phonetics and Speech Communication
University of Munich, Schellingstr. 3, 80799 Munich, Germany
{reubold, al-x}@phonetik.uni-muenchen.de

Abstract

The most obvious difference between recordings of German words and non-sense words (logatomes) within the BITS project was an audibly noticeable difference in pitch: diphones obtained from logatomes seemed to have a higher pitch than those from words. We proved this by measuring the pitches of diphones concatenated to sentences, and by comparing the pitches of actually uttered logatomes with pitches of words. The results show, that the prompted material has influence on pitch values, i.e. that the design of corpora is a non-trivial problem; besides this, logatomes revealed higher pitch values.

We refer to Ellis’ and Young’s model of human language processing and speech production to support our hypothesis, that a higher mental workload during articulation of logatomes results in a rise of pitch. The question arises whether the usage of logatomes for obtaining high quality diphone synthesis is appropriate.

1. Introduction

This paper describes whether the use of logatomes (non-sense words, that comply with phonotactical requirements of a given language) in recordings for obtaining speech material (e.g. diphones) for concatenative speech synthesis could be critical. Our main assumption is: speakers produce logatomes with a higher pitch than they do with words. This assumption is due to observations after a casting of speakers had been carried out in order to find suitable voices for concatenative speech synthesis (using diphones) within the BITS Project. Because of the advantages (see next section) of logatome corpora for embedding diphones or other units needed for high quality concatenative speech synthesis, their disadvantages have not been widely discussed yet, although it is known from literature, that logatome processing is much different from word processing (e.g. see [1]), rising mental workload, and, on the other hand, higher mental workload is known to enhance speakers’ pitch (e.g. see [2], [3] or [4]). These two effects, when combined with each other, could explain our observation.

2. Material and methods

2.1. The BITS speech synthesis corpus

"BAS Infrastructures for Speech Processing (BITS) is an initiative for the creation of a scientifically and technologically well-founded platform for resources of spoken German language which will be opened to every user on the Internet"[5]. One of its purposes is “planning, annotation and segmentation of an universally available corpus with professional speakers for concatenative speech synthesis”. For obtaining a unit-selection corpus, phonetically rich sentences were recorded, for obtaining a diphone-based corpus, logatomes complying with the phonotactical requirements of German were recorded (compare e.g.[6]). The use of logatomes is the procedure often practiced for obtaining diphone corpora, as logatomes can be freely designed to fulfill equal phonetic conditions.

2.2. Casting

In order to find suitable voices casting recordings had been carried out, using a small subset of the logatome corpus, which had been chosen to get the diphones needed for concatenation and therefore very simple synthesis of three German target sentences: “Heute ist schönes Frühlingswetter!” /h0Y3@QistS2:n@sfrY3:1lNSvEt6/, “Wer muss noch Schule machen?” /f:6mU3nOxSu:1Qx:baI/nmmax@u/ and “Der herrliche Pate versteht sich als Pol der ganzen Familie.” /de:6hERIS@pa:t@1Et6Ste:tsICQalspo:lde:6gants@nfami:l/j@/.

2.2.1. Logatome corpus

The BITS logatome corpus contains logatomes of 2-4 syllables; the prescribed default accentual pattern looked as follows: the main “word”-accent should be on the last syllable, the syllable(s) containing the diphone or parts of it should be unaccented. The latter prescription remains with the few exceptions, that are needed to obtain utterance-final “diphones” (for example: /mIU:PAUSE/ needed for synthesis of utterance-final German “machen”/mmax@u/ or English “then”/D/En/); these exceptional logatomes only consist of two syllables, and the first one is accent-bearing. All diphones, if not adjacent to silence are embedded into an articulatory neutral context, i.e., consonants are adjacent to an contextual /@/ or /ʔ/, vowels are adjacent to an contextual /ʊ/ or /ʌ/, like in /QadIsad’aU/, where the diphone is /ls/.

As the corpus is made for a reading task, prompts were made that could be shown on a computer monitor. To do this, the requirements of German for grapheme-phoneme/phoneme-grapheme conversion were complied, e.g. /QadIsad’aU/ is converted into ”ADISSADAU”.

2.2.2. German words corpus

16 of the speakers also uttered “real” German words containing the same diphones. This had been done in order to find out, whether the carrying unit (logatome vs. word) influences the quality of the recorded voice.

These words should have had at least similar phonetic conditions as the logatomes have had, i.e., as they could not show comparable articulatory neutrality, the diphone’s position should have been on not accent bearing syllables at least. In order to assure this, it was tried to find suitable words.

The difficulties in doing so are the main disadvantage of the
use of words, but they seemed to be manageable when doing it for quite a small test corpus. The test words had 2-5 syllables. Many of the words were compounds. As many of these are quite unusual, reading was simplified by using hyphenation (which was also done with some of the logatomes). With most of the words the first syllable is accent-bearing. If the accentuation pattern was ambiguous, the speakers were told to stress the first syllable.

2.2.3. Experience of the speakers

Three of six female speakers were trained speakers, but one of the non-trained females was also an experienced professional speaker. Five out of ten male speakers were trained professionals.

2.2.4. Recordings

When the two test corpora had been finished, 16 speakers had to read them in one recording session, first the words, then the logatomes. Most, but not all of the speakers read the sentences mentioned above at the end of their recording as reference speech. Three channels were recorded: the signals of a close talk microphone ("Beyerdynamic NEM 192"), a large membrane condenser microphone and a laryngograph. For further details of the high quality recording procedure see [7].

Although all three channels can be used for synthesis, the close talk channel is used in the default case. The same channel was used to measure the pitch values, as it will be described later in this paper, as it turned out that the differences between the channels’ pitch values were too small to be of any importance.

2.3. Annotation and concatenation of the casting material

The two phones building the target diphone within every logatome or word were segmented and annotated by trained phoneticians. The obtained diphones were used for synthesis in quite a simple way: Every used phone was cut at 40% of its duration, so that the last 60% of the diphones first phone and the first 40% of its second phone build up the concatenable units of synthesis. These units were smoothed to zero amplitude near their external boundaries and simply concatenated. All this is done without any other manipulation by using a Perl script especially made for BITS.

2.4. Audible differences in pitch with logatome- vs. word-diphone synthesized speech

The three sentences mentioned above, synthesized out of diphones obtained from logatomes, were used for searching for the best suitable voices for concatenative synthesis. This was done by a group of phoneticians listening to the synthesized speech, evaluating its quality in certain terms (for example intelligibility, naturalness). A by-product of this evaluation resulted from comparing the sentences made of logatome or word material: The word speech sounded as if produced with quite deep pitch (according to the pitch range known from recordings of the same sentences by the same speakers), but with most of the speakers, the logatome speech sounded as if produced with an audible much higher, nearly unnatural high pitch. These two kinds of synthesized sentences (among 27 other recordings without word speech and 2 recordings of children not used in this paper) can be listened to under “goal and auditive samples” at "http://www.phonetik.uni-muenchen.de/Forschung/BITS/index.html" [5]. There can also be found a list of the 90 used logatomes/words.

2.5. Measurements of F0

The question arose whether this perception of higher pitch in logatome diphones is physically measurable and statistically significant over the whole group of speakers and not induced by the special behavior of some of the speakers, i.e., whether it is a real logatome vs. word effect. To answer this, F0 measurements using the software PRAAT were done. The “accurate autocorrelation algorithm” was chosen. For men, a range of 50-300 Hz, for women, a range of 75-600 Hz was used. The obtained frequency values were later converted to mel values. In a first step, the median F0 values of the single sentences were measured, to get F0 within the diphones: so 2 ± 3 values for each speaker could be obtained: three for logatome speech and three for word speech. Median values were used as outlying F0 values should not damage the measurements.

It is of substantial interest, whether the logatomes, representing one type of units and the words representing a second type of units can be differentiated by means of global pitch. So, for every logatome or word the median F0 value was measured, so that the resulting numbers were 2 ± 90 values per speaker.

2.6. Dimension reduction

For the 2 ± 3 F0 values obtained from the synthesized sentences, it was shown by an ANOVA, that only 1.16% of variance of these values is due to the variable "sentence", so that these 2 ± 3 values could be reduced later to 2 ± 1 per speaker.

To get an idea what were the differences between the measurements of whole sentences and the single logatomes or words, the latter F0 values were reduced as well. As there are 90 logatome-word-pairs and 16 speakers, we had 16 ± 2 ± 90 values. These values were reduced to 16 ± 2 ± 1 values, i.e. the 90 values of logatomes or words were reduced to one value (the median value), to simplify the plotting of the results and to make the results of both methods of measurement comparable.

3. Results

After the data reductions, the measurements of the synthesized sentences give us one value per logatome speech or word speech and speaker. The difference of these two values per speaker gives us the approximated average difference within one speaker’s logatome- or word-diphones, as these diphones were the components of synthesis. It is an approximation, as a small number of diphones had been used in more than only one of the three sentences.

![Figure 1: Men’s diphone F0 values from sentences.](image1)

It is not surprising, due to anatomic differences, that men’s
vs. women’s F0 values show great differences, so it is better to divide these two groups due to statistical reasons.

As can be seen in figures 1 and 2, it is obvious, that the diphones embedded into logatomes were produced with a higher pitch than the diphones embedded into words.

The next two figures (3 and 4) show the results of measurements of whole logatomes or words (after the data reduction described above), as this is essential to evaluate the latter results without possible effects caused by potential faults in the design of the two corpora. The logatome vs. word effect still seems to be obvious, although the difference values are smaller now.

For each of the 90 logatome-word pairs a t-test was carried out to compare the mean values of the 16 speakers’ logatome

or word values: it could be shown, that 52 logatome-word-pairs were statistically significantly different, with positive difference values (logatome value - word value), 37 pairs were not significantly different, and only one pair showed significance at negative difference values. So, the majority of the logatome values is significantly higher than the word values.

Of course, there is a great variability within one speaker’s difference values. So, figure 5 shows us this variability and shows that the values for some speakers are not just chance.

For the male speakers 2, 6, 7, 8 and 9 the lion’s share of values is positive, for 4 of the 6 female speakers (here: no. 13-16) as well. But no speaker’s bulk of values is entirely negative - so, there is a tendency, that speakers really produce logatomes with higher pitch than they do with words. For a greater richness of details see [8] (in German), available under “http://www.phonetik.uni-muenchen.de/Lehre/Magisterarbeiten/UlrichReubold.pdf”.

4. Discussion

4.1. Corpus effects vs. logatome-word effects

We assume two effects, cooperating in enhancing the logatomes’ pitches’ values. As can be seen above in the results, there is evidence that most speakers produced logatomes in a higher pitch than words, and the differences of F0 values within diphones obtained from logatomes vs. those obtained from words show an even greater effect. This gives us an idea of an impact caused by the two corpora. This impact could obviously not be avoided despite the design of both corpora, that were assumed to provide similar phonetic conditions in terms of non-accentuation of the diphone containing syllables. It seems, that the position of these syllables within the logatomes or words does play an important role, as well as the prescribed accentual patterns. In [8], some details of these corpus problems are depicted.

In this paper, we concentrate on the still remaining logatome-word effect. Why did most of the speakers in the majority of utterances produce logatomes with a higher pitch than they did with words?
4.2. Hypothesis

Our hypothesis is, that the use of logatomes can lead to an elevation of pitch to the higher regions of a speaker’s pitch range due to the speaker’s unfamiliarity to non-sense words without entry in his/her mental lexicon, i.e. logatomes. This unfamiliarity causes a higher mental workload due to the missing automation of language processing and speech production. This mental workload, we suppose, is a stressor to the speaker, and this stress leads to a higher tension within the speaker’s muscles - and therefore to a higher rate of oscillations of the vocal cords, resulting in a higher pitch.

4.3. Language processing and speech production

We refer to Ellis’ and Young’s modular model of human language processing and speech production [1] to explain our thoughts: It is known, that automatic processing is done in a parallel way, i.e., processes like reading and speaking known words can be done very quickly without conscious control. This is the fact, when a known word is processed along the route visual analysis (where letters are recognized) - visual input lexicon (where written word-forms are recognized) - semantic system (its output is a word-meaning correlation) - speech output lexicon (containing all learned phonetic plans for words) - phoneme level (where the plans for transforming concepts into articulatory movements are stored). This seems to be quite a long route, but along this route, tasks can be processed in parallel in several modules at the same time, as there is no necessity for conscious control.

A shorter, but much harder and therefore longer taking route is the one used for unknown or non-sense words: visual analysis - grapheme-phoneme-conversion - phoneme level.

Here, processing has to be done step by step (i.e.: serial instead of parallel processing), as conscious control is necessary (in contrary to automatic processing). This conscious control, which is not involved in “normal” reading, enhances the speaker’s mental workload.

4.4. High mental task and its effects on pitch

It is known from many experiments, in which the mental task is enhanced by some given simultaneous tasks like, e.g., calculating numbers during car driving at high velocity (see e.g.,[3] or [4]), that enhanced mental task can heighten pitch. We assume a similar influence for our data.

Also many other conditions leading to speakers’ stress are reported to heighten pitches even in a more obvious way than the latter described stressors (see, e.g., [2]).

5. Conclusion

This paper shows, that some speakers produce logatomes in a heightened pitch compared to their word productions. According to our hypothesis, this could be due to an enhanced mental workload while a speaker is reading logatomes. This hypothesis must be tested in the near future.

It was argued, that the output of speech synthesis with diphones obtained from logatomes sounds quite different from word material concatenation. Besides pitch, future work must show, whether other features of logatome speech are affected as well. The question arises whether the recording of logatomes is the best way for obtaining units of speech needed for high quality concatenative speech synthesis. Future work is needed, to answer the question, whether speech synthesis generated with diphones obtained from logatomes could be implicitly felt as alert speech and therefore rated “unpleasant”.

6. Acknowledgments

Acknowledgments to Uwe Reichel, Olga I. Dioubina and Christiane Hofbauer.

7. References