Analysis of knowledge requirements for speech and text alignment problem

Bartosz Kalińczuk

September 25, 2013

Required input

- prepared audio model -> any similar
- limited knowledge of graphemes to phonemes conversions
- knowledge about alphabet and punctuation marks

- Graphemes to phonemes conversion
- Converting text to HMM
- Viterbi algorithm

English audio model

- "Doktor Piotr"
  After **38** seconds and after **62** words it incorrectly assigned 3 seconds to word “części” and it never recovered.
- “Boże Narodzenie”
  It has gone wrong after **257** seconds and **503** words on word “czarownicach”.

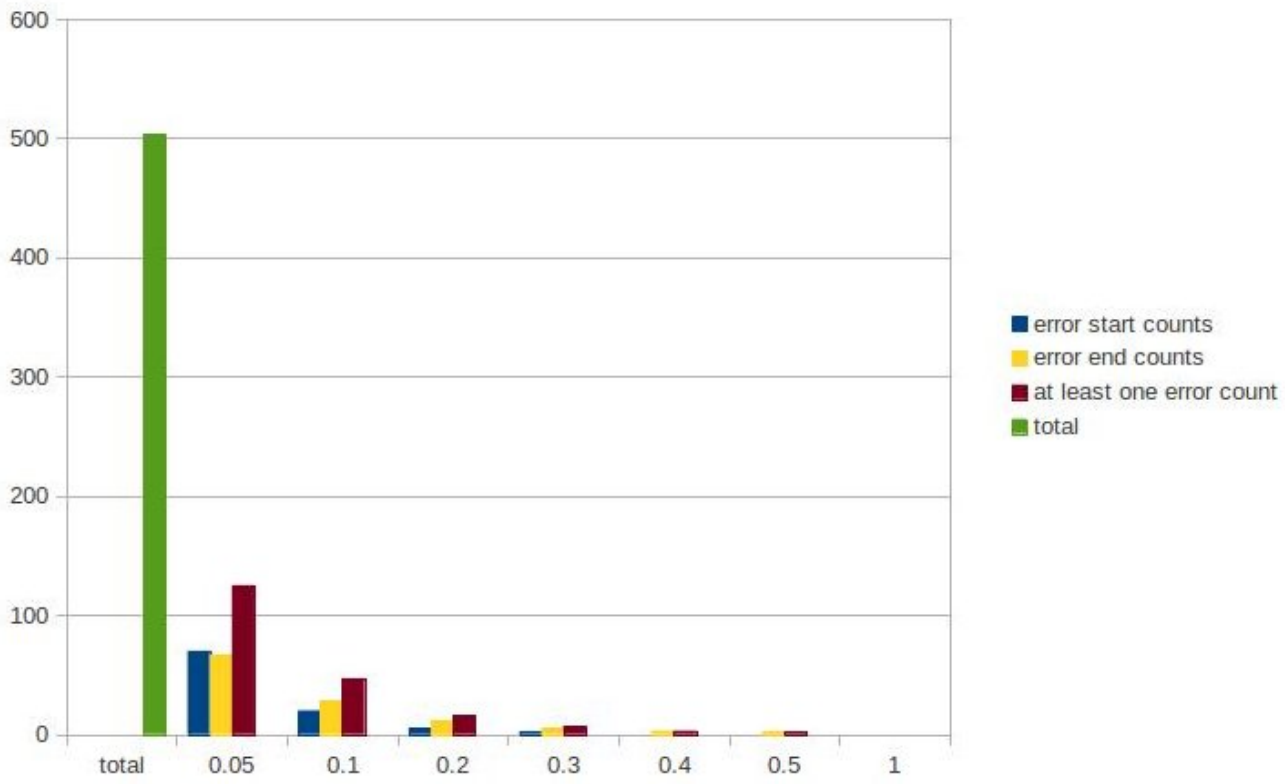
English audio model

The statistics for "Boże Narodzenie" however shows, that before it goes bad it actually aligns first **503** words quite nicely:

- Maximum difference (start or end): **0.559s**
- Maximum difference (start or end), if label was to short at one end: **0.559s**
- Average difference (start or end): **0.032s**

English audio model

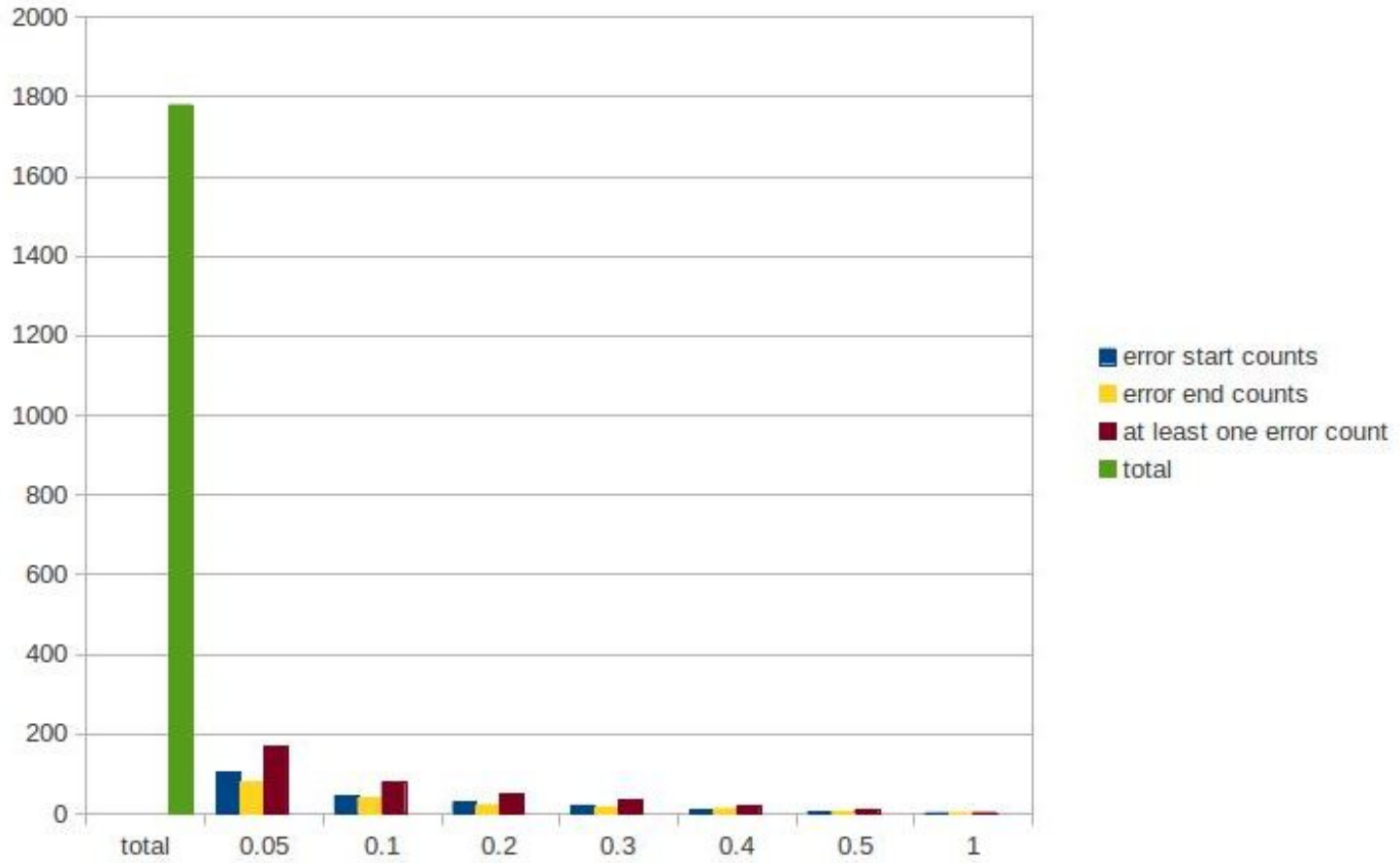
Error counts depending on time thresholds:



Russian audio model - "Boże Narodzenie" statistics

- Total number of words: **1779**
- Maximum difference (start or end): **2.451s**
- Maximum difference (start or end), if label was to short at one end: **0.543s**
- Average difference (start or end): **0.016s**

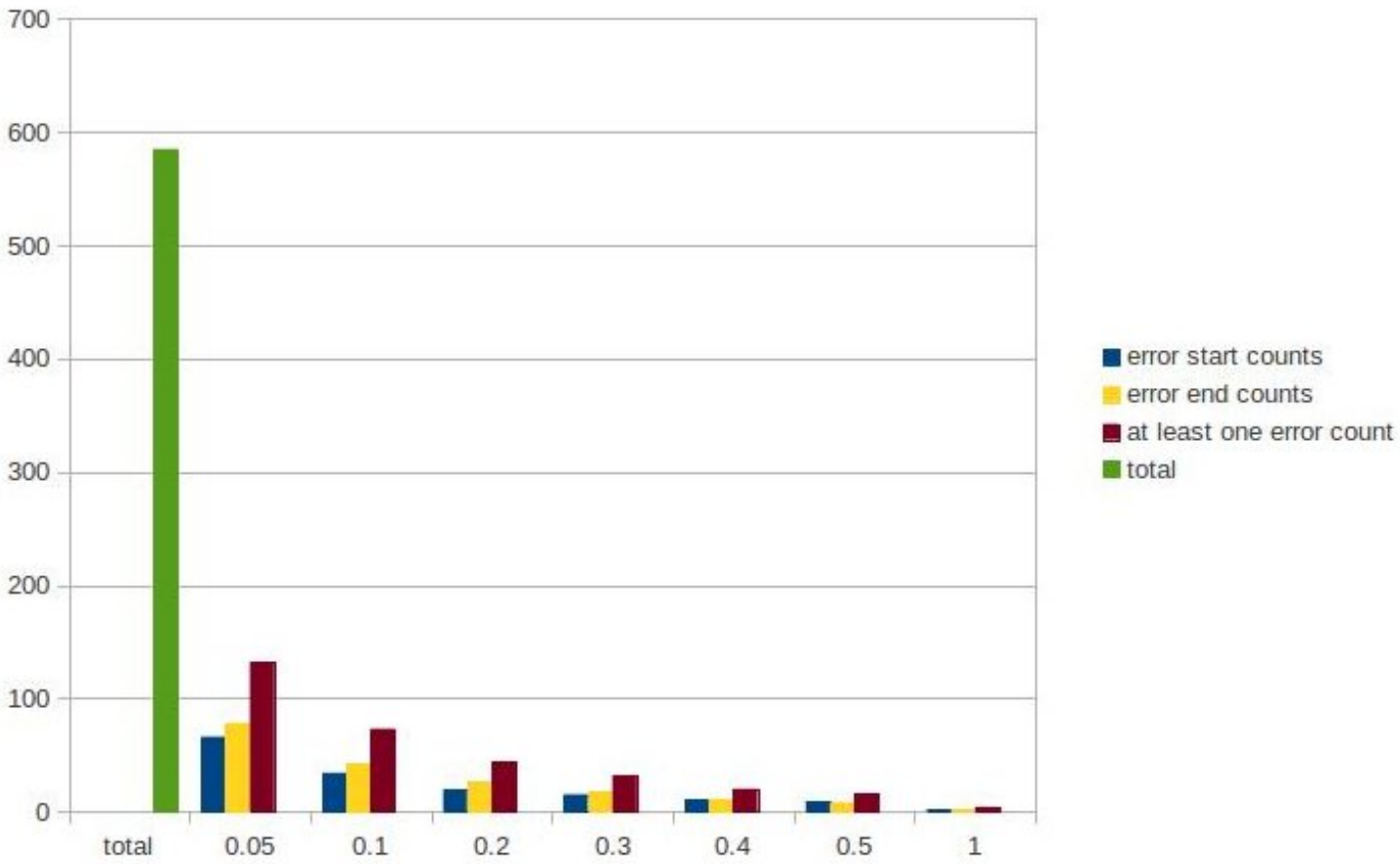
Russian audio model - "Boże Narodzenie" error counts



Russian audio model - "Doktor Piotr" sample statistics

- Total number of words **585**
- Maximum difference (start or end): **1.354s**
- Maximum difference (start or end), if label was to short at one end: **0.534s**
- Average difference (start or end): **0.033s**

Russian audio model - "Doktor Piotr" sample error counts



Testing sample

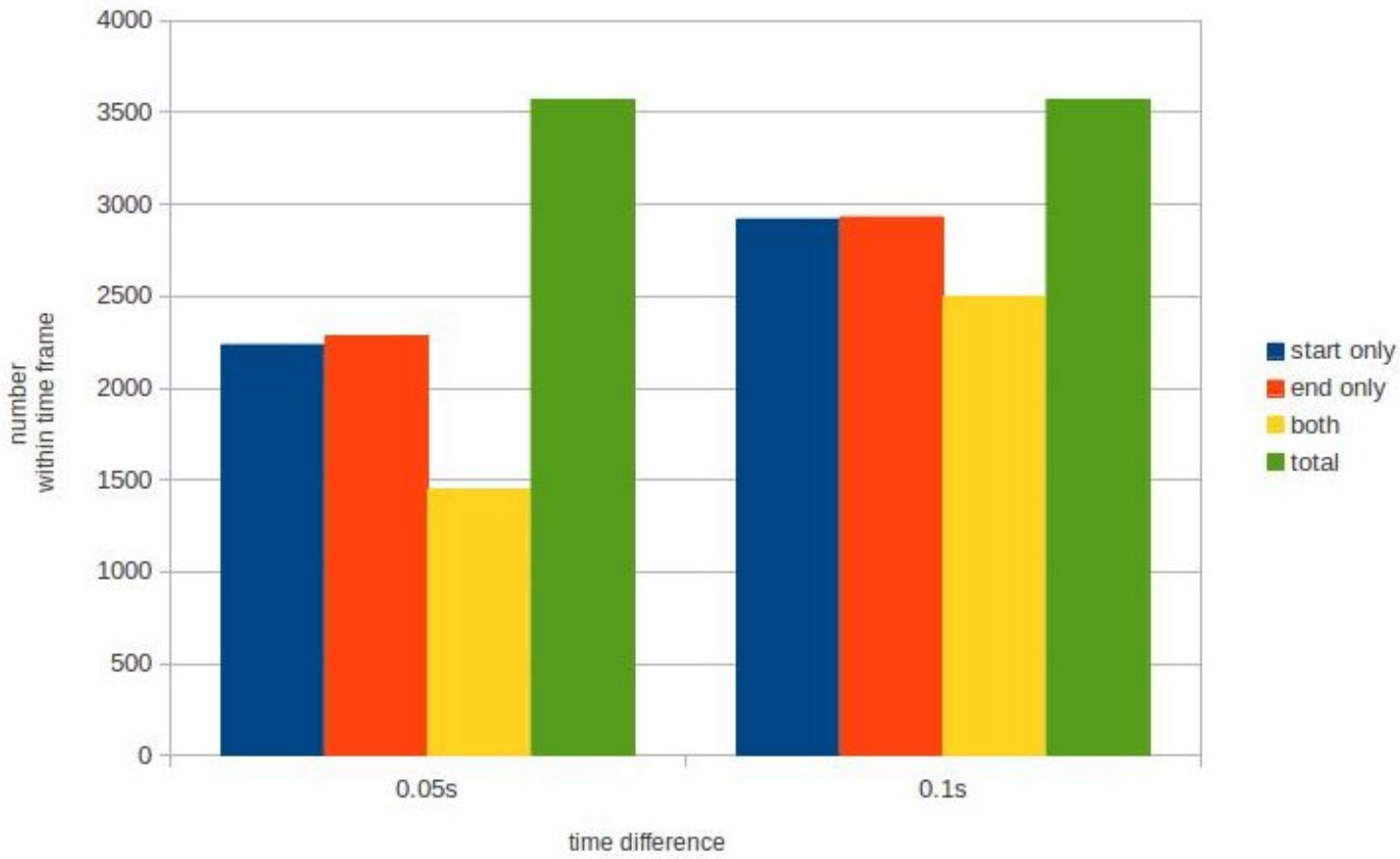
Tests are performed using Corpora corpus, which contains audio recordings of digits, names and some unusual sentences for tens of different speakers, and all of these recordings are tagged with phonemes.
Merged recording contains a 7 minutes and 26 second of audio data, a total of **843** spoken words and **3611** phoneme labels.

Statistics of phoneme alignment with Russian audio model

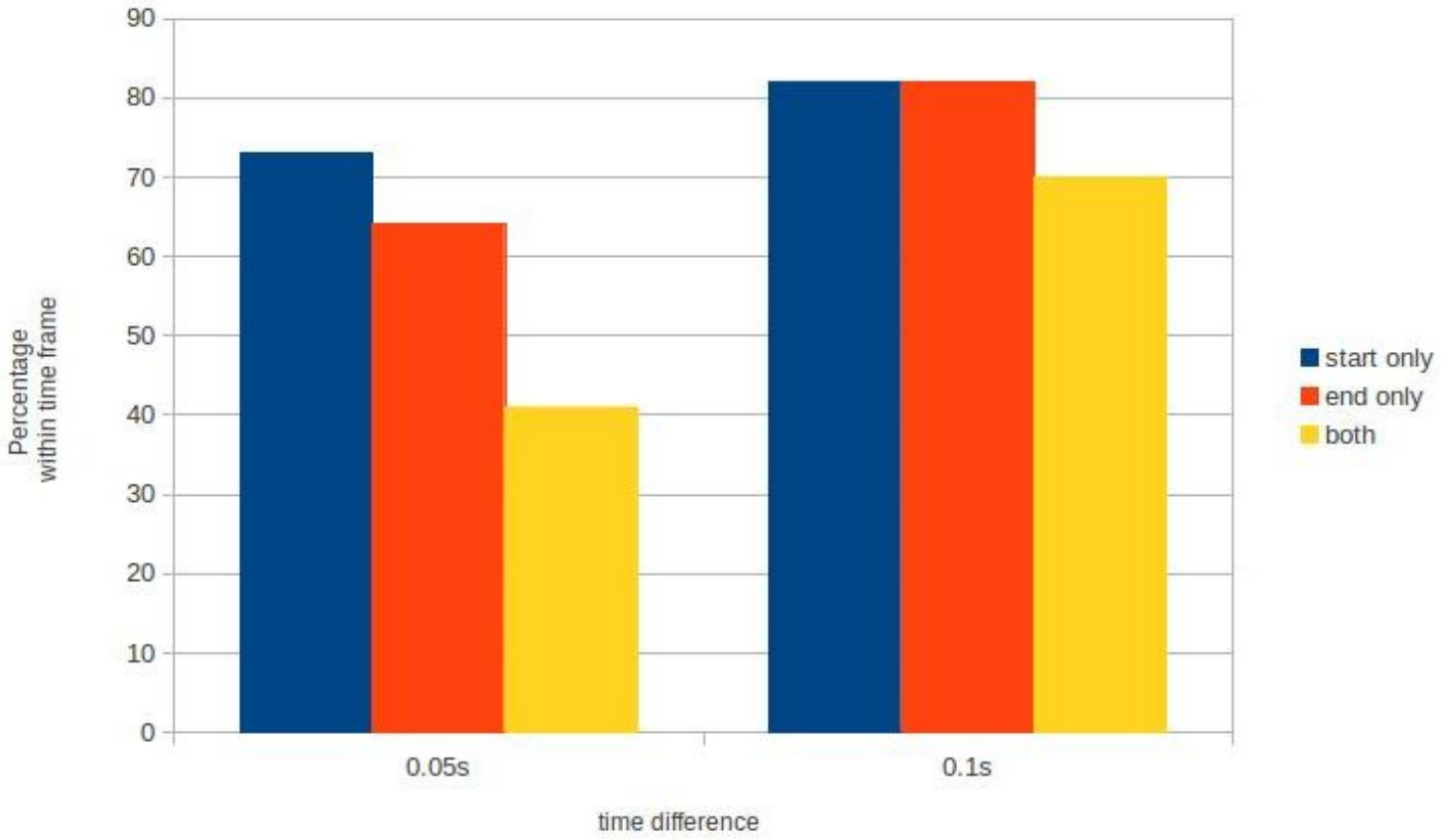
The test was A resulting match contained **3570** of pairs.

- Average start difference: **0.0571s**
- Average end difference: **0.0574s**
- Maximum time difference:**0.516s**

Error counts using Russian audio model

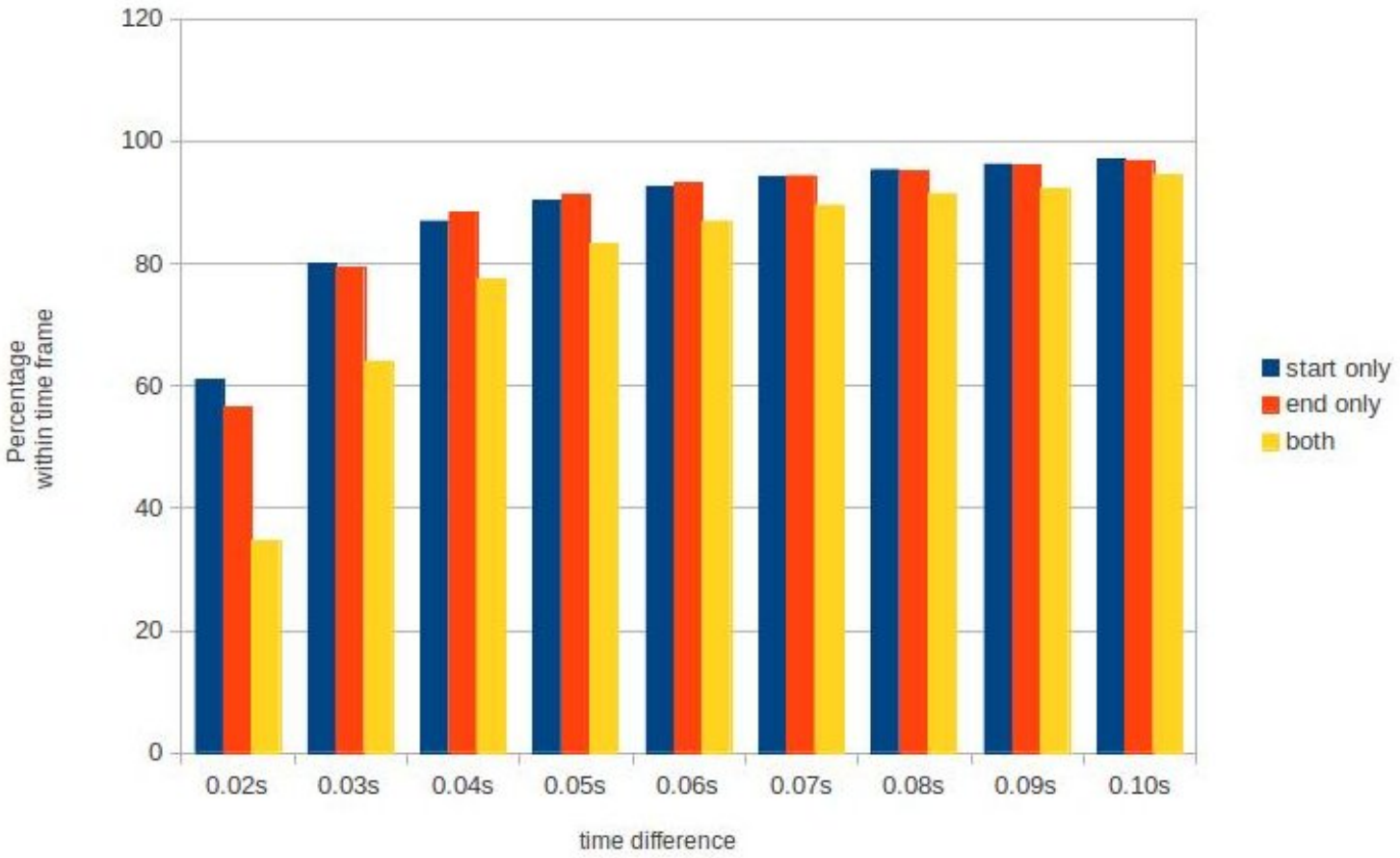


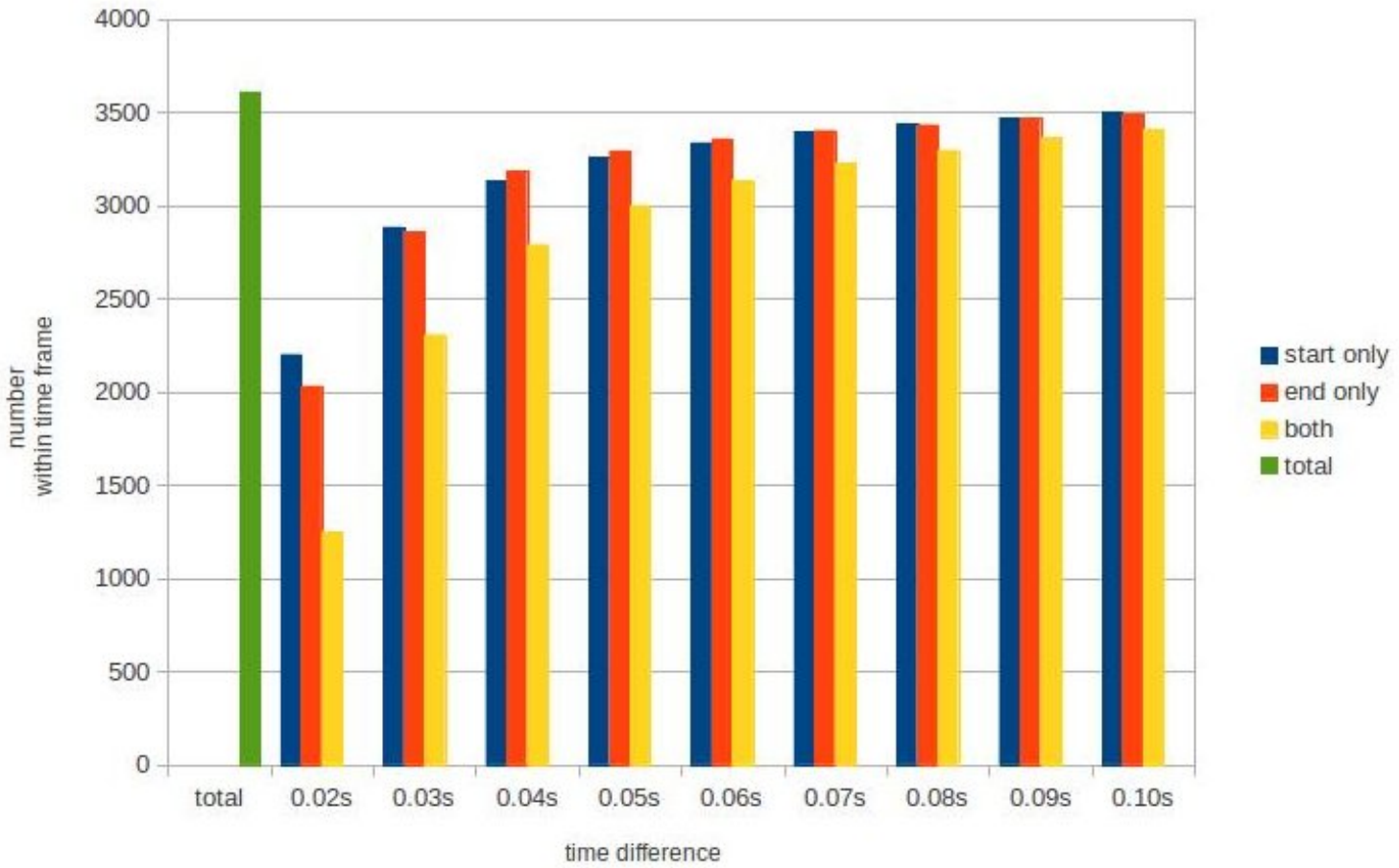
Error counts (percentage) using Russian audio model



Matched phonemes contained **3611** pairs.

- Average start difference: **0.02402s**
- Average end difference: **0.02439s**
- Maximum time difference:**0.5131s**

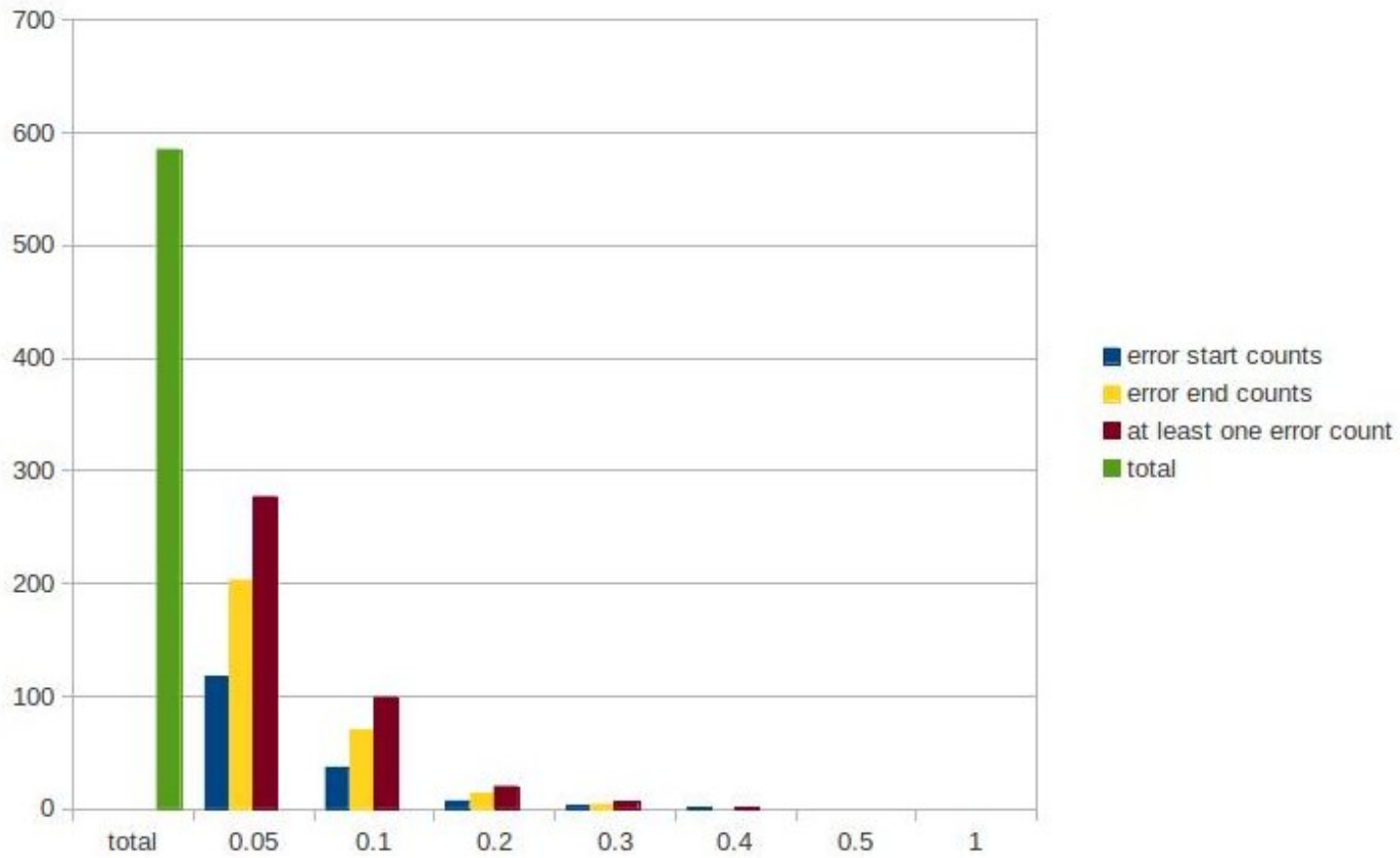

120
100
80
60
40
20
0
Percentage within time frame
0.02s
0.03s
0.04s
0.05s
0.06s
0.07s
0.08s
0.09s
0.10s
time difference
start only
end only
both


4000
3500
3000
2500
2000
1500
1000
500
0
number within time frame
total
0.02s
0.03s
0.04s
0.05s
0.06s
0.07s
0.08s
0.09s
0.10s
time difference
start only
end only
both
total

Statistics - sample of "Doktor Piotr"

- Total number of words **585**
- Maximum difference (start or end): **0.422s**
- Maximum difference (start or end), if label was to short at one end: **0.371s**
- Average difference (start or end): **0.044s**

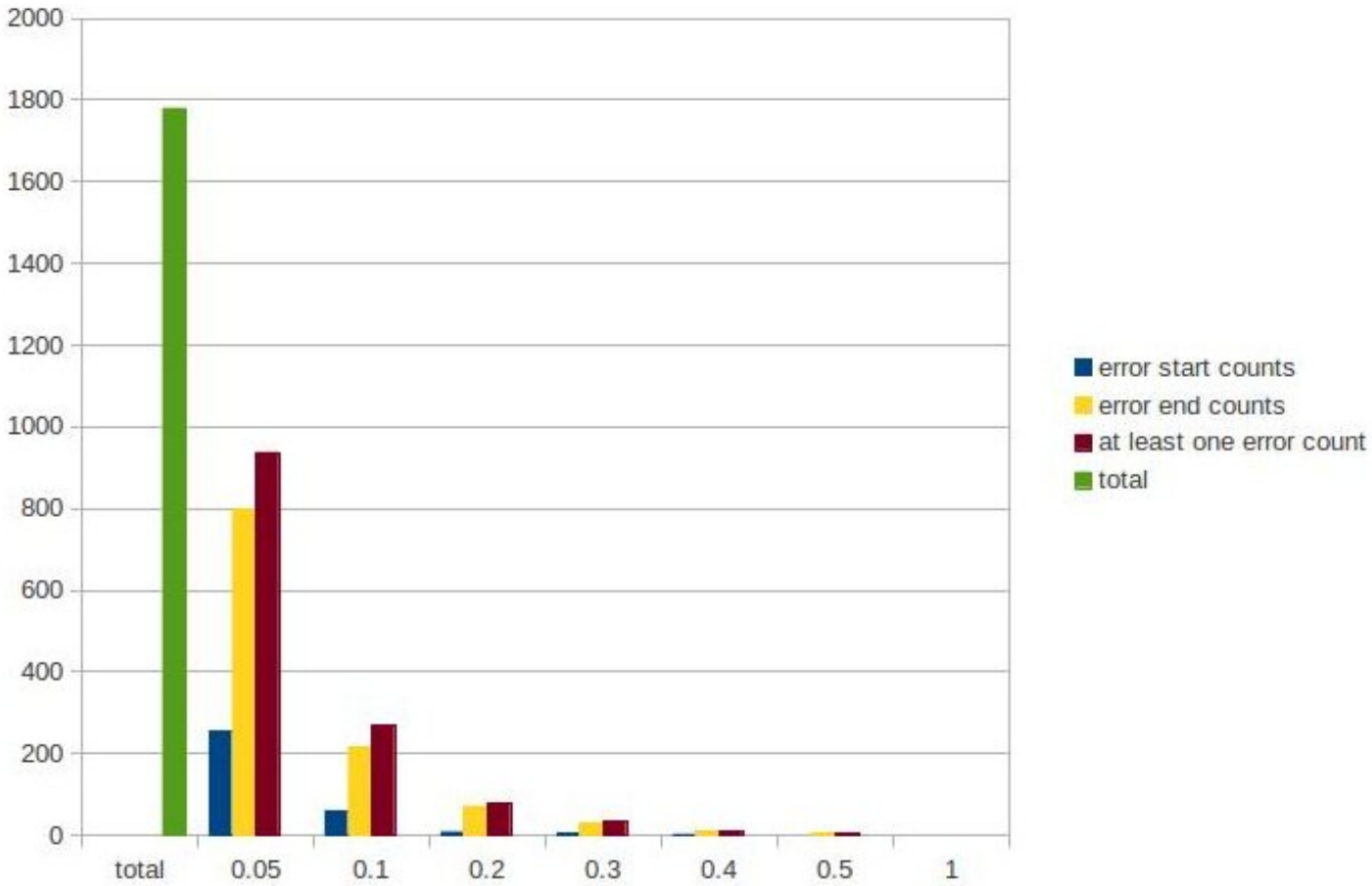
Error counts - sample of "Doktor Piotr"



Statistics - “Boże Narodzenie”

- Total number of words **1779**
- Maximum difference (start or end): **0.606s**
- Maximum difference (start or end), if label was to short at one end: **0.605s**
- Average difference (start or end): **0.046s**

Error counts - “Boże Narodzenie”



Sample synthesized texts

- "W czasie suszy szosa sucha"

Sample synthesized texts

- "W czasie suszy szosa sucha"
- "Za górami za lasami znajduję się wysoka wieża strzeżona przez smoka"

Sample synthesized texts

- "W czasie suszy szosa sucha"
- "Za górami za lasami znajduję się wysoka wieża strzeżona przez smoka"
- "Litwo Ojczyzno moja ty jesteś jak zdrowie Ile cię trzeba cenić ten tylko się dowie Kto cię stracił Dziś piękność twą w całej ozdobie Widzę i opisuję bo tęsknię po tobie"

Sample synthesized texts

- "W czasie suszy szosa sucha"
- "Za górami za lasami znajduję się wysoka wieża strzeżona przez smoka"
- "Litwo Ojczyzno moja ty jesteś jak zdrowie Ile cię trzeba cenić ten tylko się dowie Kto cię stracił Dziś piękność twą w całej ozdobie Widzę i opisuję bo tęsknię po tobie"
- "Chrząszcz brzmi w trzcinie w Szczebrzeszynie W szczękach chrząszcza trzeszczy miąższ Czcza szczypawka czka w Szczecinie"

Sample synthesized texts

- "W czasie suszy szosa sucha"
- "Za górami za lasami znajduję się wysoka wieża strzeżona przez smoka"
- "Litwo Ojczyzno moja ty jesteś jak zdrowie Ile cię trzeba cenić ten tylko się dowie Kto cię stracił Dziś piękność twą w całej ozdobie Widzę i opisuję bo tęsknię po tobie"
- "Chrząszcz brzmi w trzcinie w Szczebrzeszynie W szczękach chrząszcza trzeszczy miąższ Czcza szczypawka czka w Szczecinie"
- "Rosja przedwojenna była wymarzoną areną dorobku dla ludzi tego typu zwłaszcza pochodzących z Królestwa"

Conclusions

Future works