1918 TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY

INSTITUTE OF CYBERNETICS

Phone Duration Modeling Using Clustering of Rich Contexts

Tanel Alumäe, Rena Nemoto
Institute of Cybernetics at Tallinn University of Technology, Estonia

Motivation

How to better model phone duration in speech recognition?
HMMs model phone durations using state transition probabilities.

- This results in **geometric** probability density functions (PDFs) for phone durations
- However, phone duration distributions are rather **log-normal** or **gamma**, not geometric

Goals of our model:

- To efficiently model the dependence of phone duration on various contextual factors, such as neighboring phones, phone position in a word, word position in a sentence, intra-sentence pauses
- To automatically take into account varying and possibly non-stationary **speaking rate**

Model

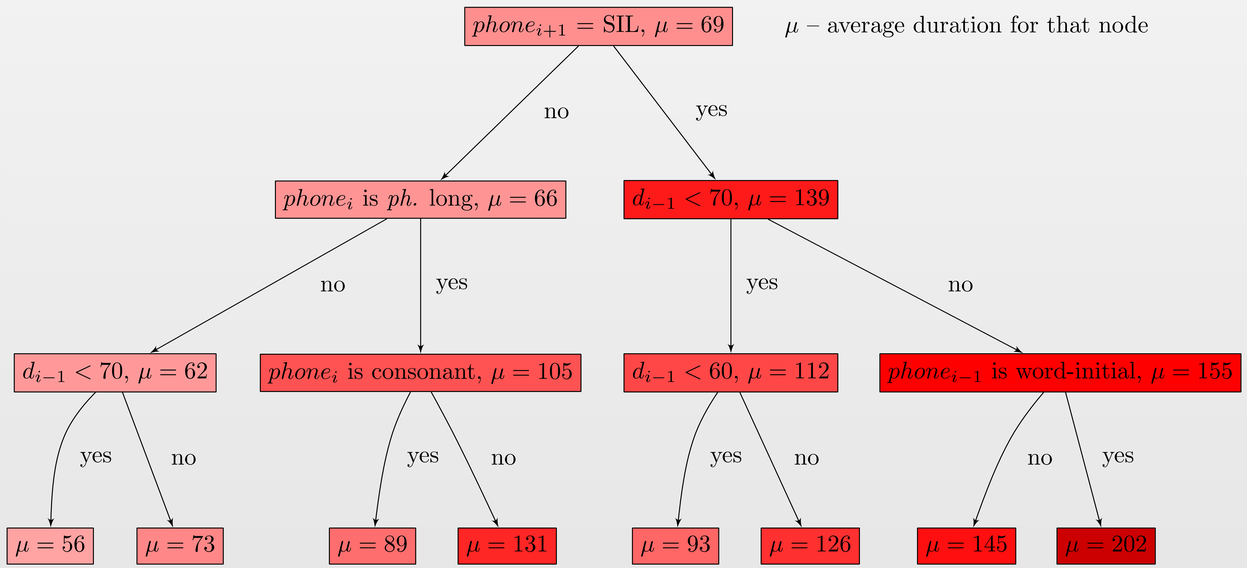
We encode phones as bag-of-features. Features are binary or discrete-valued.

- Binary features ask questions:
  - about the phone itself, for example:
    - Is the phoneme /a/?
    - Is the phoneme vowel?
    - Is the current syllable 2nd in the word?
  - about the neighboring phones (+-2), for example:
    - Is the previous phoneme /m/?
    - Is the pre-previous phone first in the utterance?
    - Is the there a pause after the current phone?
- Discrete features correspond to observed durations of the last 2 phones.

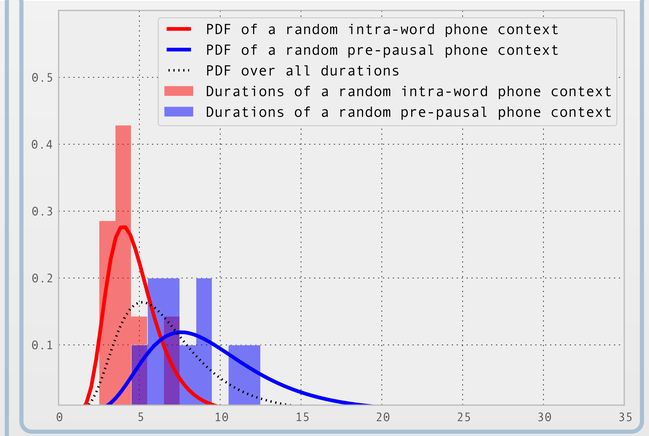
Duration model:

- **Using phonetically aligned training data**, we construct a **decision tree regression model** that minimizes predicted duration error, given the features. Each tree terminal node is required to have at least 100 samples.
- **For each tree leaf**, we estimate a **duration log-normal PDF** from the phones that ended up in this node.
- **To apply the model**, we use the decision tree on the test data in order to find a terminal node for each phone, and use corresponding log-normal PDF to estimate $P(d_i|features)$.

What does the model learn (for on Estonian)?



How different are the learned PDFs?



Which features are important?

Feature	*Importance*
$phone_{i+1}$ = silence	0.4900
$phone_i$ is phonetically long	0.3300
$duration_{i-1}$	0.1200
$phone_{i-1}$ = silence	0.0270
$phone_i$ is vowel	0.0082
$phone_i$ is utterance-ending	0.0035
$duration_{i-2}$	0.0027
$phone_{i+2}$ = silence	0.0020
$phone_i$ is utterance-starting	0.0014

Observations:

- Most useful feature asks whether **the next phone is silence**. This is related to **pre-pausal lengthening**.
- Features related to **durations of previous and pre-previous phones** are highly ranked. This confirms that the model is able to take speaking rate into account.
- Phoneme identity features are not particularly important. In fact, the highest ranked feature related to phoneme ID is for the phone /j/ which is always very short in Estonian.

Data fit *vs.* feature sets

We evaluated, how much each feature set contributes to performance of the model. We measured the likelihood of the development set duration values against the learned model. Likelihoods were converted to perplexity (PPL) for easier comparison (lower values are better).

Feature set	*#Features*	*PPL*
Phone identity, phone type	36	10.6
+ Phone phon. length, word and utterance boundary, syllable number	55	9.2
+ Features of two previous phones	157	8.5
+ Features of two next phones	259	7.9
+ Durations of two previous phones	261	7.1
+ Features and durations of four previous phones instead of two	365	7.2

Speech recognition experiments

- Estonian language
- Evaluation:
  1. Broadcast news (BNews), 2 hours for dev and test
  2. Broadcast convesations (BConv), 1 hour for both dev and test
- 140 hours of speech data for training acoustic and duration models (dictated speech, broadcast news, broadcast conversations, lectures, spontaneous speech)
- Language model: 300M words from various domains, Kneser-Ney smoothed 4-gram, separate models optimized for both evaluation domains
- Acoustic models: LDA-transformed MFCCs, ML training (RWTH-ASR), no adaptation
- Duration model learned from 4.9M phones, resulting decision tree had about 36000 leaves
- Duration models applied for N-best rescoring

Word error rate results:

	BConv		*BNews*	
	Dev	*Eval*	*Dev*	*Eval*
Baseline	37.2	32.4	13.5	13.8
+ Duration model	36.5	31.0	13.2	13.5

- 0.3-1.4% absolute (2.2-4.3% relative) improvement in WER
- Statistically significant improvements (Wilcoxon test)