TOWARDS THE QUESTION: WHY HAS SPEAKING RATE SUCH AN IMPACT ON SPEECH RECOGNITION PERFORMANCE?

R. Faltlhauser, G. Ruske, M. Thomae

Institute for Human-Machine-Communication
Technische Universität München (TUM)
Arcisstr. 21, 80290 Munich, Germany
{faltlhauser,thomae,ruske}@ei.tum.de

ABSTRACT

It has repeatedly been shown, mostly in terms of WER, that the rate of speech significantly affects speech recognition accuracy. However, the question how is not yet satisfactorily answered. In this paper we scrutinized in which way already modeling accuracy is influenced by the rate of speech. We observed the existence of a rather direct (negative) correlation between the local speech rate (LSR) and the local average HMM score (LAS). This correlation can already be found for utterances in the training database, i.e. utterances that actually were used for the parameter estimation of the acoustic phonetic models. By introducing confidence measures based on likelihood distance we verified that statistical modeling with respect to speech rate seems most accurate in slow speech segments and deteriorates already at average speaking rates. We further found that the correlation is little, yet observable, for the static features and increases with the frame range of delta(delta) features - reaching up to 0.65. The correlation persists regardless of simple monophone models or context dependent triphones. The LSR-LAS dependency can be used to predict LSR on independent test data directly from the acoustic HMM scores. In addition, LAS can be used as an indicator to assess the performance gain of rate dependent HMM models, which seems small (for fast speech) in comparison to the overall score degradation.

1. INTRODUCTION

In several studies, e.g. [1, 2, 3, 4, 5] it has been verified - in terms of word error rate (WER) - that the speech recognition performance significantly degrades at a higher rate of speech (SR). The question whether performance deteriorates really for slower speech rates is not yet satisfyingly answered: Martinez et al. [1, 2] observed strong degradation for their database, whereas e.g. Pfau [3] or Wrede [5] found highest accuracy at low rates - or at least no degradation in comparison to average speaking rate. A reason for this are possibly the databases used. The TRESVEL database collected by Martinez included, according to the authors, also slow speech artifacts which could be subsumed with the term “hyper-articulation”. Such effects are perhaps not contained in the other databases, which comprise 'standard' spontaneous or read speech.

Most studies towards speaking rate published so far, focus mainly on phonetic effects (phome duration, etc.) [1, 8], and gauge the impact by changes on the overall word error rate [2, 3, 5, 7]. The point is, measuring WER does not allow an explicit localization of the components of an ASR system (preprocessing, statistical modeling, language model (weights), lexicon), which are mostly affected and how. One of the major questions are therefore the consequences of these phonetic variations for the statistical acoustic modeling (HMM, PDFs) and its interdependence with the preprocessing stage.

In our work we focused on the influence of speaking rate already on modeling accuracy. We scrutinized how speaker independent (SI) HMM models cope with the inherent variation of speaking rate of a (spontaneous) speech corpus. For this purpose we used a local speech rate measure (LSR) to account for the exact changes of speaking rate during the utterances. We observed that LSR highly affects the local average HMM score (log, likelihood). To study this impact on the statistical HMM models we introduced the local average HMM score (LAS) supplemented by 2 local confidence measures (LC). LSR and LAS reveal a high correlation, which depends on the structure of the feature vector but is rather independent of the model type (monophones, triphones). Moreover, the dependency of LAS indicates that the capability of the statistical models degrades already at average speaking rates. In return, the correlation can be used to estimate speaking rate directly from the acoustic score. Moreover, LAS can be used as an indicator for the performance of SR dependent models [2, 3, 4, 5]. Our results show that the overall performance gain of such models - if any - seems poor in comparison to the overall loss caused by speaking rate.

2. DEFINITION OF LOCAL MEASURES

2.1. Local Speech Rate

The local speech rate (LSR) is commonly given by the quotient of the number of phonetic units within a small observation window and the length of the window. Martinez [1] for example computed the LSR in a window covering a constant number $N_W=5$ of phonemes. This results in a dynamic window length and a measurement shift of accordingly one phoneme. In contrast thereto we adopted in our experiments the strategy proposed e.g. by Pfitzinger [9], who used a window of constant length $N_w$, which can be shifted frame-wise:

$$ LSR(n) = \frac{b_{i+1} - b_i - n \cdot w_G(n) \cdot w_S(n) - b_i}{b_{i+1} - w_L(n) + w_H(n) - b_i + \sum_{j=i+1}^{i+1} (b_{j+1} - b_j)} $$

$I$ and $r$ are the first and last segment (partly) included by the window, and $b_i$ gives the left segmentation mark for the $i$-th seg-
ment. \(w_L(n) = n - \frac{N_F}{2}\) and \(w_R(n) = n + \frac{N_F}{2} - 1\) mark the window boundaries for current frame \(n\). The phonemes which are only partially overlapped by the window are considered according to the proportion to which they are covered. In case of a constant window length, the denominator sums up to \(N_F\). In our experiments we used a length of \(N_F = 100\) Frames. In the following we will refer with the term LSR to this particular implementation. In our experiments we computed the net [9] speaking rate, which only considers units actually present in the speech signal. For this reason we used a segmentation allowing for pronunciation variants - instead of a canonical segmentation. A sample of a LSR (in [Phones/s]) curve is depicted in Figure 1 (top graph), which already indicates that the LSR is highly varying during an utterance.

### 2.2. Local Average Score

In our experiments we focused on the impact of speaking rate on modeling accuracy. Simply expressed, speech frames should achieve a high probability when emitted by the according model of the phonetic unit. Modeling accuracy can therefore be indicated by the probability or, in the log-likelihood domain, by the score average. In order to study the influence of speaking rate on this variation we compute the local average score (LAS) to mark the range of the feature vector we varied the use of delta(\(\Delta\))-coefficients. For each setup individual models were

\[
\text{LAS}(n) = \frac{1}{N_F} \sum_{j=n-\frac{N_F}{2}}^{n+\frac{N_F}{2}-1} S_{m,s}(x_j) = \frac{1}{N_F} \sum_{j=n-\frac{N_F}{2}}^{n+\frac{N_F}{2}-1} \log p_{m,s}(x_j)
\]

Figure 1 shows a comparison of LSR (top graph), score (middle) and LAS (bottom) for the sample utterance 'g091a000' of the German Verbmobil spontaneous speech corpus. Already from this first example it can easily be observed that the LAS seems highly correlated with local speaking rate.

### 2.3. Local Confidence

It can be argued that the score by itself is not meaningful enough. The question is how the other competing models perform for the same rate. For this reason we defined 2 local measures of confidence (LC) which should cope with this question. The first measure, \(LC_{top}\), is given as the average of the score distances to the most rivaling (=top-1) models of the correct models. The correct model is provided by the segmentation, whereas the top-1 model is searched per frame locally among all other models.

\[
\Delta S_{top1}(j) = S_{m,s}(x_j) - \max_{m \neq m_s} S_{m,s}(x_j)
\]

\[
LC_{top1}(n) = \frac{1}{N_F} \sum_{j=n-\frac{N_F}{2}}^{n+\frac{N_F}{2}-1} \Delta S_{top1}(j)
\]

### 3. EXPERIMENTS

#### 3.1. LSR and LAS

Our sample utterance indicated the existence of a strong correlation between the local speaking rate and the local score. In order to verify this observation we evaluated LSR and LAS for the German Verbmobil spontaneous speech corpus. As first experimental setup we used: monophone HMMs, 8.5k Gaussians, and features consisting of 12 MFCCs, energy, zero crossing rate, together with 1st and 2nd derivatives. Figure 2 shows a (downsampled) scatterplot of LAS vs. LSR, together with a linear and a logarithmic regression curve. Pause-, non-speech as well as the leading and trailing \(\frac{N_F}{2}\) frames were excluded. The distribution shows a remarkably high correlation coefficient of \(\rho_{LAS} = p(LSR, LAS) = -0.64\), resulting in a conspicuous decline of scores towards fast speech. We see this behavior as critical (especially) in relation to the balance with the probabilities of the language model (and LM weight), which is commonly adjusted only ‘on average’.

#### Dependence on feature vector

In order to examine the reasons for the speaking rate induced effect we computed the correlation coefficients \(\rho_{feas}\) for each utterance in the training database individually. Since the degradation was expected to be dependent on the structure and ‘range’ of the feature vector we varied the use of delta(\(\Delta\))-coefficients. For each setup individual models were

![Figure 1. Comparison of LSR, −Score and −LAS. (Please note, LAS is inverted to −LAS only here for better visualization.)](image-url)
trained. Figure 3 depicts histograms of the calculated correlation coefficients. Utterances shorter than 2 seconds were excluded, since the utterances should be long enough to ensure a wide dynamic range of speaking rate within the utterance.

Figure 3 (top left graph) indicates a median correlation coefficient of \( r_{\text{med}} = -0.16 \), which states that already the static features are slightly affected by speaking rate. However, the extension by \( \Delta \)-coefficients (top, right) significantly increased the median correlation to \( r_{\text{med}} = -0.37 \). Comparing these histograms with the lower 2 graphs, which comprise also \( \Delta \Delta \)-features (both \( r_{\text{med}} = -0.65 \)), show that the influence of speaking rate continuously increases with the 'range' of the dynamic parts of the feature vector. A reason for this can be seen in the fixed \( \Delta \)-computation grid which can reach more than 1 neighboring phoneme at high rates. The growing context sensitivity accordingly increases the variability of (\( \Delta \cdot \)) features, and in the end, makes it more difficult to robustly represent all possible patterns.

**Context dependent models:** The above experiments were conducted with monophone models. It can be argued that triphone models show a better coverage of context dependent phenomena and thus, of context induced speaking rate dependent diversity. For this reason we conducted the above experiment with triphone models (~35k Gaussians, CART based state tying), which used also static+\( \Delta + \Delta \)-coefficients. Figure 4 shows the resulting histogram of the utterance-wise correlation coefficients. The histogram with \( r_{\text{med}} = -0.55 \) clearly states that even the model-

with more sophisticated topology and more Gaussian densities cannot overcome the speaking rate dependent behavior, although the performance is slightly more independent compared to monophone models.

![Figure 2](image1.png)

**Fig. 2.** LSR and LAS show a correlation coefficient \( r_{\text{LSR}} = -0.64 \) for the training corpus.

![Figure 4](image2.png)

**Fig. 4.** Distribution of correlation coefficients of the training utterances using triphone models.

**3.2. Confidence Measures**

\( LC_{\text{topi}} \), as well as \( LC_{\text{mean}} \) are computed locally for a window of \( N_F \) frames, equivalent to the calculation of LAS. In Figure 5 scatterplots of both measures are depicted. Both plots indicate a decrease in score distance between the correct and the rivaling models towards fast speech, although not as distinct as for LAS. However, LAS as well as both confidence measures evince highest accuracy of the statistical modeling in regions with low speaking rate. This raises the question, whether speaking rate normalization techniques [6, 11], which normalize slow speech to average speaking rate (or phoneme length), are truly sensible - since length reduction would mean a waste of more reliable information.

![Figure 5](image3.png)

**Fig. 5.** \( LC_{\text{topi}} \) and \( LC_{\text{mean}} \) for the training data. \( r_{LC_{\text{topi}}} = -0.11 \) and \( r_{LC_{\text{mean}}} = -0.25 \).

**3.3. Speaking Rate Prediction**

Some authors have already proposed to estimate speaking rate from acoustic features. E.g. in [2] or [10] parallel SR dependent models were used to classify the rate category. An approach based on a specific energy measure was proposed by Morgan [7]. Our precedent experiments indicate that speaking rate can be measured directly from the acoustic scores of the recognizer HMMs. Thus, we used the linear regression curve determined on the training data (Fig. 2) to predict LSR from the LAS of the test data. The upper graph in Figure 6 shows the (downsampled) global scatterplot of the LSR vs. LAS for the test database together with the reference regression and the regression of the training data. LAS was determined on the top-1 recognizer hypothesis, whereas the reference LSR was taken from the correct phonetic transcription. Both regressions show basically the same behavior, with only a slight vertical offset which can be related to the difference of training and test data. The lower graph in Figure 6 depicts the resulting scatterplot for the estimated vs. the reference LSR having a corre-
lation coefficient of 0.00.

Fig. 6. Top graph shows LAS vs. LSR distribution ($\rho_{LAS} = -0.60$) of test data together with regression curves for test and of training data. Bottom graph depicts estimated (via LAS) vs. reference LSR.

3.4. LAS Rating of SR Dependent Models

A number of authors, e.g. [3, 4, 5] have presented the training of speaking rate dependent models as a method for SR compensation. The idea is to use the appropriate model set in the recognition phase depending on the actual SR category. Most of the authors categorized the training utterances using a global (per utterance) SR measure. In order to examine the potential of the turn-wise approach we trained (MAP) models for 3 classes: slow, average and fast speech. Since LSR highly varies during an utterance, we propose the training of models based on the LSR criterion, which allows per-frame data separation. Therefore we trained further 'slow' and 'fast' models categorized by LSR. Like in Figure 6 (test data) we extracted from the LSR-LAS distribution of each model a representative linear regression curve, which are depicted in Figure 7 (bottom: absolute regression; top: relative to rate independent models). The slopes of the resulting curves basically point out 2 facts: the respective rate class model is indeed able as intended to outperform the other rate classes in the according class range (top graph). So in the mid-range (~10-15 Phones/s) the 'average' models dominate, whereas at higher rates the 'fast' models and at lower rates the 'slow' models score better. In addition, our proposed 'LSR models' are able to outperform their 'global' counterparts. However, it is quite obvious, that the slight improvement - if any - cannot overcome the overall slope of the score degradation which is considerably steeper (bottom graph).

Fig. 7. Bottom Graph: LAS-Regression curves on test data for rate dependent models. Top: LAS gain over SR independent models.

4. DISCUSSION

In this paper we could show that the acoustic HMM score rather directly reflects the rate of speech. By introducing the local average score (LAS) we could verify a high negative correlation between the local speaking rate (LSR) and the short-time (local) average score, represented by LAS. The correlation is little, but observable for the static features and increases with the use of $\Delta$- and $\Delta^2$-features - reaching up to $-0.65$ on the training data and $-0.00$ on test data. The choice of a fixed delta computation grid seems to contribute to this problem. Moreover, the dependency is rather persistent for monophone as well as context dependent triphone models. LAS verifies best scoring of slow speech segments, a fact that makes rate normalization techniques questionable, which shorten slow speech segments to average speech rate. By means of a regression curve, the correlation can be used to predict local speech rate directly from the acoustic score. In addition LAS can be used to appraise the modeling ability of rate dependent models, which have shown to offer only little performance gain in correcting the overall score slope. Our next step will therefore be the study of SR normalization techniques, which seem to have more potential.

5. REFERENCES