A FAST AND RELIABLE RATE OF SPEECH DETECTOR

Jan P. Verhasselt [1] *and Jean-Pierre Martens* [2]

ELIS, University of Ghent
St.-Pietersnieuwstraat 41
B-9000 Gent (Belgium)
verhaslt/martens@elis.rug.ac.be

ABSTRACT

In this paper, we present a new rate of speech (ROS) detector that operates independently of the recognition process. This detector is evaluated on the TIMIT corpus and positioned with respect to other ROS detectors. The ROS estimate is subsequently used to compensate for the effects of unusual speech rates on continuous speech recognition. We report on results obtained with two ROS compensation techniques on a speaker independent acoustic phonetic decoding task.

1. INTRODUCTION

The performance of automatic speech recognizers typically degrades for unusually fast or slow speakers [1]. It has been shown that compensation techniques can reduce the errors for fast speech in HMM [2,3,4,5] as well as in hybrid HMM/MLP [6] recognition systems. However, these techniques require a reliable ROS detector. In the first part of this paper, we present and evaluate a new ROS detector, which can be used prior, during or after the recognition search. Subsequently, the advantages and drawbacks of each of these approaches are analyzed and the proposed detector is positioned with respect to other ROS detectors. Finally, we address a number of ROS compensation techniques focusing on the influence of ROS on phone durations and on spectral features.

2. ROS DETECTOR

By rate of speech, we mean the rate at which individual speech units are uttered. Reported ROS measures differ in the choice of the speech unit that is used in the calculation. It has been argued [2,3] that phone rate is more suited than syllable or word rate. By normalizing the phone durations with respect to the phone specific expected durations [4,5] and variances [2], a normalized phone rate can be obtained that is very effective in differentiating utterance rates. However, this requires phonetic segmentation and classification information that is bound to be provided by the recognition process. Therefore, *normalized* phone rates can only be calculated during [4] or after [2,5] the recognition search.

In this paper, we show that the *unnormalized* phone rate, defined as the number of phones per second, can be estimated by a ROS detector operating independently of the recognition process. Recently, it has been reported [3,6] that such a ROS measure too, even though it is unnormalized, is valuable for compensating the effects of fast speech.

The ROS estimate (ROS^e) is obtained by accumulating the phone boundary evidences in a certain interval, and by subsequently dividing the result by the duration of that interval. The phone boundary evidences are provided by a small Multi-Layer Perceptron (MLP) that was trained to estimate for each hypothesised boundary the posterior probability that it is a phonetic segment boundary. A boundary can be hypothesised on each frame (which is the approach explored in this paper), or on a limited number of time instants which were selected by a presegmentation algorithm. The proposed detector estimates the number of phone boundaries and thus the number of phones per second.

The MLP has one output, 11 hidden nodes and 50 inputs. The inputs consist of the auditory spectrum in the vicinity of the boundary and some change functions measuring spectral and total energy changes. The training examples were extracted from 160 sentences. For each 10ms frame boundary, a training example is generated. The training targets were obtained from the hand segmentation that comes with the TIMIT corpus. If the frame boundary corresponds with a phone boundary, then the target is one, otherwise it is zero. If no hand segmentation is available, a forced alignment would be required in order to obtain the phonetic segmentation.

The length of the interval used in the calculation should be short enough to account for changes in rate of speech during the utterance, while long enough to contain enough phones, as to yield a rate that is not too much affected by the phonetic content. Since the TIMIT utterances are fairly short, the ROS was computed over a whole sentence. In order to prevent silences from disturbing the ROS estimate, non-speech segments are discarded from the calculation.

A scatter plot of the actual rate of speech (ROS^a), as derived from the hand segmentation, versus ROS^e is shown in figure 1. The solid line shows the best linear fit through the data. The dotted line shows the unbiased predictor. The observed bias is due to imperfections in the boundary probability estimates that are provided by the MLP. The sign and magnitude of the bias shows an arbitrary dependency

[1] Aspirant N.F.W.O. - Belgacom
[2] Research Associate of the National Fund for Scientific Research

on the choice of the MLP inputs, the network size and the training parameters. However, the ROS estimate is monotonously related to the actual ROS, and therefore an improved estimate can be obtained by regression. In the experiments reported below, we used a linear regression (the solid line in figure 1), which was determined on the sentences of the TIMIT training corpus that were not used for the MLP training. Higher order regressions did not yield a substantial improvement.

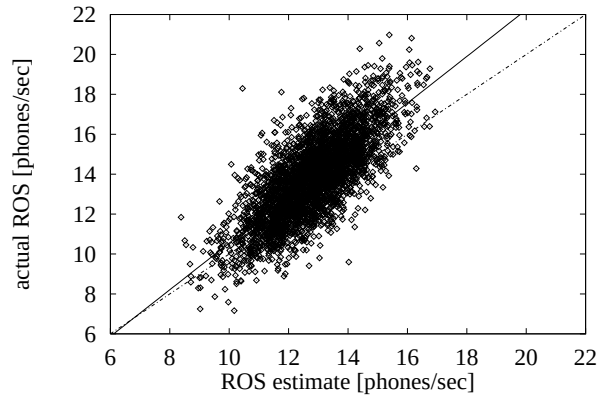

Figure 1: Scatter plot of the actual ROS versus the ROS measure. The solid line shows the best linear fit through the points. The dotted line shows the unbiased predictor

The error between the predicted and the actual ROS approximates a zero-mean Gaussian distribution with a standard deviation of 1.36 phones/sec (1.38 phones/sec without regression), whereas the standard deviation of the actual ROS is 2.03 phones/sec. Figure 2 shows a histogram of the relative prediction error, defined as:

$$\varepsilon_R = 100 * \frac{ROS^e - ROS^a}{ROS^a} \quad (1)$$

The standard deviation of the relative prediction error is 9.9% (9.0% without regression). This has to be compared with a standard deviation of 16% when the mean ROS (13.83 phones/sec) is used as the ROS 'estimate'.

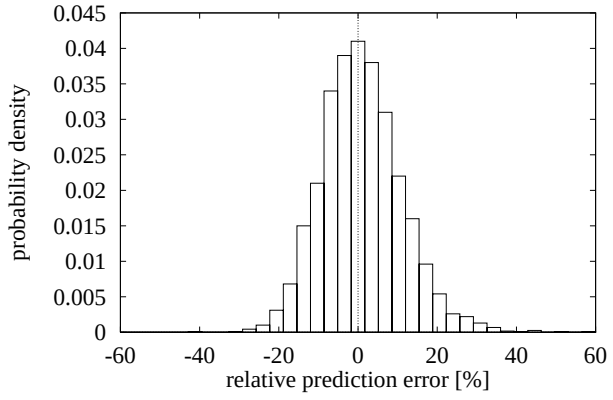

Figure 2: Histogram of the relative prediction error of the ROS detector.

For comparison, we also took the number of phones per second in the best phone string hypothesized by our phone recognizer as a ROS estimate. The standard deviation of the absolute error was now 1.35 phones/sec, corresponding to a relative error of 9.6% (8.9% without regression). Apparently, this alternative ROS estimate is not better than the one proposed above. Moreover, it can only be calculated during or after the recognition process.

3. WHERE TO USE THIS ROS DETECTOR?

The major advantage of the proposed ROS detector is that it does not require a recognition process. Therefore, the algorithm is simple and fast. The computational cost is limited to the cost of detecting silences and computing boundary evidences. This is an obvious advantage for applications requiring nothing more than a ROS estimate, e.g. in speech therapy.

In speech recognition, the ROS detector can be used prior, during or after the recognition search. In the next paragraphs, we will analyze the advantages and drawbacks of each of these approaches. First of all, it is important to note that the proposed detector is most valuable for applications where the ROS has to be determined in the time interval that has to be recognized. If the ROS of the previous time interval were a good estimate for the ROS in the present time interval (in other words: if the ROS shows no abrupt changes), then it would be more appropriate to calculate a normalized ROS measure from the recognition system's transcript and time-alignment of the previous time interval. However, we observed on the TIMIT corpus, that the standard deviation of the prediction error is 2.25 phones/sec if the actual ROS of the previous sentence of the same speaker is used as a prediction for the present sentence. This figure is significantly larger than the 1.36 phones/sec one obtains by using our ROS detector on the present sentence.

In the experiments reported in section 4, the ROS estimate was calculated *prior* to the recognition. The ROS is assumed constant during a sentence, but it can change arbitrarily from one sentence to the next. This prior computation has the advantage that, during the recognition, duration and/or acoustic models (and for word recognition also word pronunciation and language models) can be used which are adapted to the ROS of the sentence. On the other hand, this technique has the disadvantage that the recognition can only start after the completion of the utterance. This approach was also used in the multi-pass search algorithm reported in [4]. The mean syllabic duration was measured on an entire first recognition hypothesis and subsequently used to adapt the subword unit durational characteristics which are used in a second recognition pass. Obviously, the first recognition search is computationally more expensive than our proposed algorithm.

If the time delay introduced by the previous approach is unacceptable, the ROS can be calculated as a running average *during* the recognition process, such that improved estimates are obtained as a larger fraction of the sentence is uttered. The performance of this approach will inevitably depend on the quality of the initial estimate, especially in the beginning of a sentence. Typical choices for the initial estimate are the statistical mean of the ROS and the final esti-

mate of the previous sentence. This approach was used in [4], where a speaking rate factor is deduced for each recognition hypothesis by means of a Kalman filter. In that case, the ROS calculation is integrated in the search and at each time instant, a normalized ROS estimate is calculated for each recognition hypotheses.

In [2,5], the ROS is used the rescore an N-best list. A normalized ROS estimate is calculated on each of the N recognition hypotheses and corresponding time-alignments. With these estimates, the duration models are adapted and the obtained duration likelihoods are combined with the recognizer scores to reorder the N-best list. Since the ROS is now determined *after* the recognition process, the estimate cannot be used to adapt the acoustic models in order to compensate for ROS effects in the feature space. Moreover, this approach requires an N-best search, which is more demanding than a single-best search. Although our ROS detector could be used after the N-best search, it is our impression that, in this case, it makes more sense to derive a normalized ROS estimate from the recognition transcripts and corresponding time-alignments.

4. COMPENSATION OF ROS EFFECTS

In this section, we describe two attempts to compensate for the effects of unusual ROS. These compensation techniques are evaluated on a speaker independent acoustic phonetic decoding task, with a Context-Independent Connectionist Stochastic Segment Model [7] recognizer, using a unigram phone language model. Figure **??** shows the phone recognition performance of the unadapted system. The best second order regression of the data shows a small, but clear dependency on the ROS. The recognizer was trained on eight sentences (5 sx + 3 si) of 429 speakers from the TIMIT corpus. The reported results were obtained on the remaining 33 training speakers.

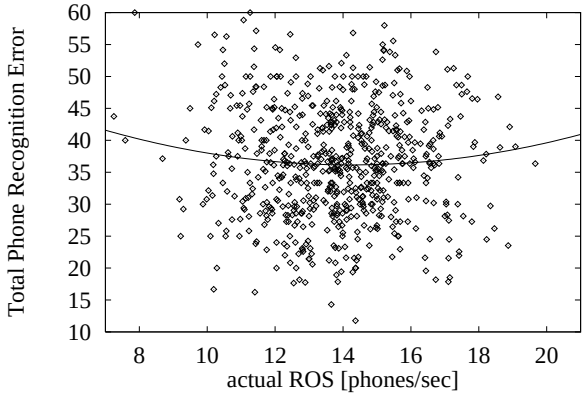

Figure 3: Total Phone Recognition Error as function of the actual ROS. The line shows the second order regression.

The system comprises a presegmentation module which generates a set $\overline{b}$ of candidate phonetic segment boundaries. A phonetic segment boundary is defined as a boundary between the acoustic realizations of subsequent phones. The segments enclosed by two consecutive candidate phonetic boundaries are called 'initial segments'. Candidate phonetic segments are built by concatenating up to five consecutive initial segments. A Viterbi search examines several candidate phonetic segmentations (characterized by a sequence of segment boundaries $\overline{s} \subset \overline{b}$) and phone sequences $\overline{u}$ of the same length as $\overline{s}$, and maximizes the joint probability of $(\overline{s}, \overline{u})$, given the acoustic evidence $\overline{x}$ and eventually the ROS of the sentence. For this purpose, the search requires the posterior probabilities given by equation **??**:

$$P(s_i = b_{n+j}, u_i = U_m | s_{i-1} = b_n, j, d, X, ROS) \qquad (2)$$

In this expression, $s_i = b_{n+j}$ means that the i-th phonetic boundary is b_{n+j}, $u_i = U_m$ means that the phone U_m (from an inventory of phones) was uttered in this segment and d is the segment duration. The vector X represents the acoustic evidence (spectrum, total energy, voicing,...) in the segment and its close surroundings. This expression can be factorized as:

$$P(s_i = b_{n+j} | s_{i-1} = b_n, j, d, X, ROS) * P(u_i = U_m | S, j, d, X, ROS) \qquad (3)$$

In this equation, we substituted the combination of $s_i = b_{n+j}$ and $s_{i-1} = b_n$ by S, which means that the segment is a phonetic one. The first factor, which we call the segmentation probability, is estimated by a MLP that is trained on all candidate phonetic segments starting on a phonetic boundary. The second factor, which we call the classification probability, is estimated by a MLP that is trained on phonetic segments only.

4.1. Modification of Acoustic Models

The dependency on X, d and j of the probabilities in equation **??** is modeled by giving them as inputs to the MLP's. The ROS dependency could be modeled in the same way. However, for the experiments reported in this paper, we followed another approach. The training sentences were split into 3 groups (slow, average, fast), based on the ROS of the sentence. The partition is done so that each group contains approximately the same number of sentences. First, a general segmentation and a general classification MLP were trained on all the data. Starting from these two networks, three ROS-specific MLP pairs were trained, one on each ROS partition, until maximum performance on a cross-validation set was obtained.

These networks were subsequently embedded in different phone recognition systems. Four systems were evaluated:

- **System-A:** Uses general MLP's (no ROS effect compensation).
- **System-B:** Uses the selected ROS-specific MLP pair.
- **System-C:** Uses a ROS-independent average (weights 1/3) of the ROS-specific MLP pairs.
- **System-D:** Uses a ROS-dependent weighting of the ROS-specific MLP pairs.

The total phone recognition error rates in table 1 indicate that, although the differences are small, the ROS-specific systems (B and D) consistently outperform the ROS-unspecific ones (A and C). Furthermore, the estimated ROS performs nearly as good as the actual ROS.

	System-A	System-B	System-C	System-D
Actual ROS	38.1%	37.7%	37.9%	37.3%
ROS estimate	”	37.9%	”	37.4%

Table 1: Adaptation of acoustic models to ROS. Phone recognition results: Total Error Rate.

4.2. Modification of Duration Models

In this section, we focus on the ROS dependency of the duration models. In order to isolate this effect, we have rewritten the classification probability in equation **??** as:

$$\frac{P(u_i = U_m|S, j, X, ROS)P(d|u_i = U_m, S, j, X, ROS)}{\sum_k P(u_i = U_k|S, j, X, ROS)P(d|u_i = U_k, S, j, X, ROS)} \quad (4)$$

The classification MLP was trained on all the data (ROS-unspecific), but the duration was not provided as an input to the network. Furthermore, once the phone identity is available, the dependency of the probability of d on X and j is neglected, so that the duration models are simplified to $P(d|u_i = U_m, S, ROS)$. This formulation allows us to model the segment duration explicitely, instead of using the implicit modeling of d by the MLP's as in section 4.1. For each phone, three smoothed histogram representations of the duration were constructed, one for each ROS partition.

To illustrate the differences between partitions, figure 3 shows the duration histograms for the vowel /ih/. The solid lines show the distributions obtained using the actual ROS for partitioning the data. The dotted line shows the corresponding distributions when the ROS prediction was used. The data indicate that our ROS estimate does not introduce severe aberrations.

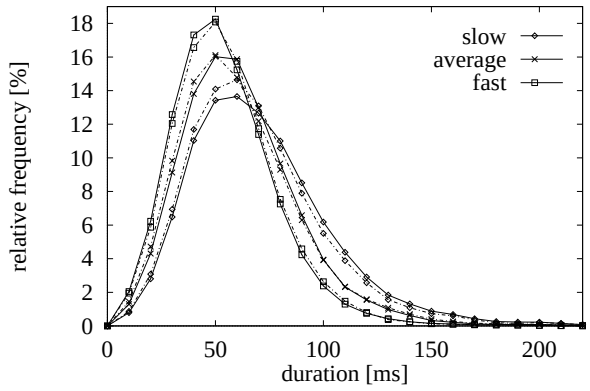

Figure 4: Smoothed histogram of the durations of /ih/ found in the slowest, average and fastest sentences

We have integrated the ROS dependent duration models in our phone recognizer. During the recognition, the phone duration histograms of the corresponding ROS partition are selected. The error rates in table 2 are lower than in table 1 because larger segmentation and classification MLP's were used for this experiment. Again, four systems were evaluated:

System-A: Does not use a duration model.
System-B: Uses ROS-unspecific duration models.
System-C: Uses ROS-specific duration models, using ROS estimate.
System-D: Uses ROS-specific duration models, using actual ROS.

System-A	System-B	System-C	System-D
36.6%	36.2%	36.0%	35.9%

Table 2: Adaptation of duration models to ROS. Phone recognition results: Total Error Rate.

The ROS estimate yields basically the same improvement in phone recognition as the actual ROS. However, these improvements are too small to be significant.

5. CONCLUSION

In this paper, a new rate of speech (ROS) detector, based on phone boundary probabilities provided by a Multi-Layer Perceptron, is presented. The detector offers a fast and reliable prediction of the phone rate, and accomplishes this without requiring a speech recognition search. When used to compensate the effects of ROS in continuous speech recognition, the ROS estimate performs nearly as good as the actual ROS that is derived from the hand segmentation. The reported compensation techniques result in a small but consistent improvement of the recognition performance.

6. REFERENCES

1. Pallet, D.S., Fiscus, J.G., Fisher, W.M., Garofolo, J.S., Lund, B.A., and Przybocki, M.A. “1993 WSJ-CSR Benchmark Test Results,” *ARPA's Spoken Language Systems Technology Workshop*, Princeton, New Jersey, March 1994.
2. Jones, M, and Woodland P.C. “Using relative duration in large vocabulary speech recognition,” *Procs EUROSPEECH*, Vol. 1, 311-314, 1993.
3. Siegler, M.A., Stern, R.M. “On the effects of speech rate in large vocabulary speech recognition systems,” *Procs ICASSP*, Vol. 1, 612-615, 1995.
4. Suaudeau, N., Andreé-Obrecht, R. “An efficient combination of acoustic and supra-segmental informations in a speech recognition system,” *Procs ICASSP*, Vol. 1, 65-68, 1994.
5. Anastasakos, A., Schwartz, R., Shu, H. “Duration modeling in large vocabulary speech recognition,” *Procs ICASSP*, Vol. 1, 628-631, 1995.
6. Mirghafori, N., Fosler, E., Morgan, N. “Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes,” *Procs EUROSPEECH*, Vol 1, 491-494, 1995.
7. Martens, J.-P. “A connectionist approach to continuous speech recognition,” *Procs FORWISS/CRIM ESPRIT Workshop*, 26-33, Munich, 1994.