An investigation of modelling aspects for rate-dependent speech recognition

Britta Wrede, Gernot A. Fink, Gerhard Sagerer

Bielefeld University, Faculty of Technology
P.O. Box 100131, Germany
{bwrede|gernot|sagerer}@techfak.uni-bielefeld.de

Abstract

For the modelling of speech rate variation in speech recognition many approaches have been suggested. However, the training of speech-rate dependent models has by far received most of the attention. In order to investigate problematic aspects related with the classification of the speech data which represents one of the major problems of these approaches extensive experiments were carried out on a German corpus of read speech. The results indicate that while the kind of the model-driven speech-rate measure is only of minor importance a data-driven classification of the speech data significantly improves the performance of rate-dependent models. Further results suggest a detailed modelling of speech rate based on more general models. This means that it might be possible to model speech rate adaptation by means of a transformation based on a continuous measure.

1. Introduction

The problems caused by speech rate variations in speech recognition are widely known and many approaches have been reported on how to increase the recognition rate in highly variable data. These approaches can roughly be subdivided into four categories: (1) training of rate-dependent models (2) new topologies of Hidden-Markov-Models (HMM) with output probabilities attached to the state transitions depending on speech rate (3) tuning of the language model and word penalty (4) manipulation of the features.

Among these approaches the training of rate-dependent models has by far received the most attention ([1], [2], [3], [4]). The underlying idea is simple. First a division of the training data into two to four subsets according to their speech rate is carried out. This is generally done by using the training segmentation to compute the phone or syllable rate or by using a signal based speech rate measure. Then rate-dependent models are trained with these subsets of the training data. During recognition an estimation of the speech rate determines which rate-dependent models are chosen for recognition.

However, no systematic variations of some possibly important parameters have been carried out. For example, little is known about the influence of the kind of the speech rate measure or the number of classes into which the training set is subdivided. Also, there is a lack of knowledge concerning some more technical aspects of the adaptation procedure such as the adaptation algorithm or the initialisation of the models. All these parameters might be important when seeking for a flexible approach for a continuous modelling of speech rate. Therefore, some of these aspects have been investigated thoroughly on a German corpus of read speech.

For an introduction into the manifold aspects of speech rate variation modelling a short survey of different approaches taken in speech recognition followed by some phonetic evidences concerning the acoustic correlates is given in the next section. Section 3 concentrates on the system configurations used for the present investigation while experiments and results are presented in section 4. A short discussion will be given at the end of the paper.

2. Related Work

As has already been pointed out most of the approaches for speech rate modelling are based on the training of rate-dependent models. However, different approaches focusing on other aspects will also be briefly addressed. In order to understand the underlying effects of speech rate variation better a short review of some phonetic evidences reflecting the spectral changes will follow.

Speech rate modelling In [1] rate-specific Multi Layer Perceptrons (MLP) were trained to classify speech into fast or slow speech by using the Perceptual Linear Predictive (PLP) features plus energy and the first order derivatives. This procedure yielded a correct rate classification of 73% for all phones and of 80-90% for all vowels. For recognition the MLP phonetic estimator was adapted to fast speech by re-training it for three epochs with the 5% fastest sentences of the training corpus. This adaptation procedure decreased the word error rate (WER) on the fast test utterances by 14% while it increased the WER of the slow utterances by 10%.

A quite similar approach was taken in [4] where a Gaussian classifier based on the dynamic mel frequency cepstral coefficients (MFCC) was trained to classify speech into fast, average, or slow speech. In a preliminary experiment these dynamic features proved to be affected to a large extent by speech rate while the effect on the static features as well as on the transition probabilities was significantly lower. The rate specific models reduced the WER for slow speech by 64% and for fast speech by 19%. A pre-classification of the test data into slow, average, and fast speech yielded slightly better results than the Gaussian classifier.

In [2] a MLP was used to estimate the phone boundaries from the signal to compute the phone rate over whole utterances. To compensate for speech rate effects previously trained general MLPs were adapted with three rate-specific subsets of the training data which reduced the WER by 4%.

While all these approaches share the use of a phone rate based classification of the test and training data in [3] a new measure is introduced. In order to estimate the syllable rate

This research was partly supported by the German Research Foundation (DFG) within the Graduate Program "Task-oriented Communication".

In Proc. European Conf. on Speech Communication and Technology, volume 4, pages 2527-2530, Aalborg, 2001.

directly from the signal a low-pass filter of 16 Hz is applied. After the filtering only the energy modulation remains which is assumed to be closely correlated with the syllable rate. In an adaptation step the transition probabilities were adjusted to the speech rate which yielded a reduction of the WER of about 10%.

However, these approaches all have to deal with the problem of sparse data since the division into subsets reduces the actual training data for each model to a fraction. Therefore, in [5] the number of parameters to be trained has been reduced by optimising the number of Gaussians per state. In [6] maximum a posteriori estimation (MAP) was applied which is reported to be well suited for problems with little adaptation data. An additional speaker adaptation by means of vocal tract length normalisation (VTLN) showed an additive effect on the reduction of the WER. Interestingly, the VTLN proved to be most effective on speakers with a high speaking rate.

A promising approach to avoid sparse data problems is reported in [7] where additional speech rate dependent output probabilities are attached to the transition probabilities. However, no experiments have been carried out and thus the effectiveness of this approach remains to be shown. The weight of the language model and the word penalty also seem to have a significant influence on the performance. In [8] the tuning of these weights led to a reduction of the WER of 21% for slow and 7.8% for fast speech.

A completely different approach was taken in [9]. Instead of adapting the model parameters to the changed speech signal the computation of the features from the signal was modified. In fast speech new frames were inserted by copying or interpolating new features from the actual frame which reduced the WER significantly. It could also been shown that speaker adaptation with maximum likelihood linear regression (MLLR) is additive to this effect leading to a WER reduction of 23.4% for fast speech.

In summary, the training of rate-dependent models yields reductions of the WER of up to 30%. However, most of the enhancement is due to a better modelling of the slow speech, where reductions of up to 60% are reported ([4]) while the WER of fast speech is generally reduced by less than 10%. In order to better understand the dynamics of the performance of rate dependent models on the different test sets with regard to their speech rate we carried out extensive experiments. Most of the approaches with rate-dependent models used an adaptation scheme where general models were adapted with rate specific data. However, it is questionable if such generalised models are able to capture rate specific effects well. One might argue that a rate specific training right from the beginning would model the rate specific characteristics more precisely. This question has been addressed in the following experiments. Also, in the above mentioned approaches much effort is put on the classification of the test data into rate classes by optimising the speech rate estimation. However, a more direct approach where the classification is left to the rate-dependent models might perform equally well without the effort of an explicit rate estimation. Furthermore, the high variation of speech rate suggests a more detailed approach since a distinction of two or three classes can hardly model this whole range adequately. Therefore, the question of an optimal number of rate classes has been addressed in our experiments as well.

Acoustic correlates All the measures used for the classification of the training and test data reported so far are based on the assumption that changes in the spectral domain of the signal can be predicted by the durational characteristics such as phone or syllable rate or some signal based estimates of these rates. However, this assumption is not unanimously shared in the literature concerning the acoustic correlates of speech rate or stress. Although there is evidence that fast speech causes heavy centralisation of the formant frequencies within the vowel space ([10], [11]) or an increase in coarticulation ([12], [13]) it has also been shown that this is not always the case. In fact, under certain circumstances this centralisation tendency in fast speech is compensated by faster movements of the articulators ([14]) such that for example unstressed vowels in clear speech are less centralised than stressed vowels in casual speech. Therefore, it can be argued that an additional measure of centralisation or coarticulation should be more appropriate than a rate based measure alone.

3. System Configurations

All experiments were carried out with a semi-continuous HMM system built within the ESMERALDA environment [15].

Corpus The training and test corpus SLACC (Spoken LAnguage Car Control) on which the experiments were carried out consists of read utterances recorded in a car environment with the speaker sitting in the front passenger seat at a distance of about 50 cm to the microphone. The utterances contained instructions that are likely to be used for the control of non safety-relevant functions such as radio or air-conditioning. The training set consists of 18 speakers (5 female, 13 male) with over 9,000 utterances while the test set contains 1,787 utterances from 4 speakers (1 female, 3 male).

Baseline System The codebook of the baseline system consists of 512 Gaussian mixtures with diagonal covariance matrices. The 39-dimensional feature-vector computed every 10 ms is based on 12 mel frequency cepstral coefficients (MFCC) plus the signal energy from which the first and second order derivatives are computed. The lexicon consists of 650 words that are modelled with around 1400 triphones. After a clustering-procedure 1600 different states are established.

Speech Rate Measures Three duration based and one formant based measures were computed on the segmentation of the test and training corpora based on a forced alignment using a phonotypical transcription. In order to classify the utterances into speech rate classes the means of the durations of all triphones (DUR), of all vowels (VDUR), and of all syllables (SYL) were computed over each utterance with the syllable boundaries set according to the phonotypical transcriptions.

The formant based measure was computed by means of the ESPS-formant tracker[1] using only the formant frequencies from the middle of a vowel. The measure for reduction (RED) is based on the positions of the vowel-tokens in the vowel space spanned by the first two formant frequencies. It captures the distance of a vowel-token T to the speaker-specific centre C normalised by the average centre-distance of all the tokens of this vowel. M denotes the the mean of all tokens of a vowel, hence the monophone M. In the following definition $||C - T||$ and $||C - M||$ denote the Euclidean distances of these points.

[1]The Entropic Signal Processing System (ESPS) is a commercial software package distributed by Entropic Research Laboratory, Inc.

$$RED = \frac{\|C-T\|}{\|C-M\|}$$

A ratio smaller than 1.0 means that the vowel token is reduced as compared to the monophone. Since reduction is often measured by means of the vowel space area it is not easy to find a simple reduction measure in the literature. Therefore, the above measure was chosen which is in broad accordance with [10] where reduction is measured in terms of relative frequency differences between the formants of stressed and unstressed vowels.

Rate-specific and rate-adapted models In order to compare the performances of highly specialised models to models still containing some information of more general speech data two different kinds of models were trained. For the training of the rate-specific models only rate-specific data was used. In contrast, the rate-adapted models are first trained with the whole data until an optimal performance is reached. Only then a rate-adaptive re-estimation of the model parameters is performed with the according subsets of the training data using the Baum-Welch algorithm.

4. Experiments and Results

The baseline system achieved a word accuracy of 75.0% which is used as the reference point for the following results.

Rate-specific vs rate-adapted models Table 1 shows that the rate-adapted models improve the word accuracy of the baseline system up to 76.6% which represents a reduction of the WER of 6.4%. This improvement is achieved in all rate classes similarly after only one adaptation step. Further iterations of this adaptation yielded roughly the same results. In contrast, the performance of the rate-specific models was slightly but not significantly worse than that of the baseline system.

Test	slow	avg.	fast	all
Baseline	79.9	78.5	68.4	75.0
Rate-specific	80.5	78.0	67.4	74.7
Rate-adapted	81.5	80.1	69.7	76.6

Table 1: Word accuracies of rate-specific models vs rate-adapted models (VDUR, 3 classes, best score).

Speech Rate Measures Table 2 shows the results of the rate-adapted models of different speech rate measures with three different classes. As a comparison a random pseudo-measure RAND has been introduced to test whether potential improvements are due to some artifacts related to the training and test procedure. However, the RAND condition produced the same result as the baseline system. All other measures provide an improvement of the word accuracy with the durational measures producing consistently better results.

Measure	RAND	DUR	VDUR	SYL	RED
Best score	75.1	76.4	76.6	76.6	76.1
Pre-class.	75.0	75.9	76.3	76.1	74.9

Table 2: Word accuracies of rate-adapted models with different speech rates. Pre-classified test condition vs best scoring hypotheses (3 classes).

Pre-classification of test data In most of the previously mentioned experiments much effort was put on the pre-classification of the test data by optimising the estimation of the speech rate. However, it is questionable if such a pre-classification is optimal. It could be argued that a data driven approach where the best scoring hypothesis is chosen might perform equally well. Therefore, the following two test conditions were compared.

The first condition consists of an optimal pre-classification of the test data according to the speech rates computed by the segmentation of a forced alignment on the test data. This condition is comparable to the above mentioned approaches. For the data driven approach the best scoring hypothesis generated by the different rate-adapted models is chosen.

Experiments were carried out on all four speech rate measures. The results as shown in table 2 indicate that the data driven approach consistently produces better results. This effect is increased when more rate classes are applied (table 3).

Number of Classes To investigate the influence of the number of rate classes into which the training data is divided systems with 3, 6, 9, 12, 15, and 18 rate classes were trained. Table 3 shows the word accuracies of these systems using the VDUR-measure for classification.

The best performance is achieved with six rate classes. However, despite the fact that the training data is dramatically reduced when using even more rate classes the slope of the word accuracy is amazingly flat. Even with 18 rate classes where only 500 utterances are used for adaptation an improvement of the word accuracy from 75% to over 76% can still be obtained.

# Classes	3	6	9	12	15	18
Best score	76.6	76.9	75.9	76.6	76.4	76.1
Pre-class.	76.3	75.9	74.8	74.7	74.4	73.6

Table 3: Word accuracies of rate-adapted models in relation to number of models (VDUR).

Rate-adaptation vs speaker-adaptation It could be argued that the 18 rate classes have specialised on the 18 speakers of the training since speaking rate can be highly speaker specific. In order to investigate this a speaker adaptation was performed where the data of each speaker was used to train one speaker-adapted model. However, the results in table 4 show that such speaker-adapted models actually degrade the performance from 75% to 72.3%.

Test-Speaker	000	002	009	014	all
Baseline	64.8	83.0	69.0	85.0	75.0
Rate-adapted	64.9	83.3	70.3	87.1	76.1
Speaker-adapted	60.3	78.1	65.0	85.5	72.3

Table 4: Word accuracies of rate-adapted models (VDUR) vs speaker-adapted models. (18 classes, best score)

A closer examination of the influence of the speaker adapted models on the test speakers revealed that two factors were affecting the performance severely: (1) the gender of the test and training speakers (2) the environment of the recordings i.e. the model of the car. Thus, the speaker-adapted models in this experiment showed to be highly sensitive to contextual factors whereas the rate-adapted models were able to generalise over these factors while using the same amount of adaptation data.

5. Discussion

The experiments on the SLACC corpus indicate that the speech rate measure only plays a minor role although the durational measures performed consistently better than the formant based one. This is contrary to the assumption that formant based measures should be better predictors for spectral degradation as has been pointed out in section 2. However, reduction might not be the best indicator since coarticulation has a greater distorting effect on the spectrum. The formant-based measures might therefore provide room for further improvements.

Secondly, the results show that rate-adapted models are more robust than models that are exclusively trained with rate-specific data. This means that a general basis is needed before a rate-dependent adaptation is performed. This seems to be especially the case for fast speech where the performance of the rate-specific models is particularly poor as compared to the baseline system. The results also indicate that all speech rates benefit roughly to the same degree from rate-adaptation.

The relatively stable performance of the rate-adapted models with a high number of classes is worth mentioning. Even with 18 rate classes the performance is still better than that of the baseline system. The experiments mentioned in section 2 only used between two and four rate-classes probably to avoid sparse data problems. However, the results of our experiments could indicate that a rather small amount of adaptation data is sufficient for rate adaptation. But it could also be interpreted that the increase in the precision of the rate modelling is very high but neutralised by the effect of sparse data. This suggests that a detailed modelling of speech rate might be more successful than the classification into only few speech rates.

Most interestingly the data driven approach choosing the best scoring hypothesis without any prior knowledge of the speech rate whatsoever proved to be consistently better than the pre-classification of the test data. This effect was even stronger when the number of classes increased. If this effect proves to be consistent over different corpora this might solve a major problem of the speech rate modelling approaches reported in section 2 where much effort was put on the classification of the test data. At least for our experiments it shows that the rate measures used for the classification of the training data are by far not optimal and a data driven classification could be more appropriate.

6. Conclusion

The fundamental problems addressed in this paper mainly concern the rate-based classification of speech. It could be shown that the different speech-rate measures yielded comparable performances while the unsupervised classification of the test data reached a superior performance. Therefore, a combination of both the top-down definition of the rate measure and its bottom-up optimisation by unsupervised classification might further improve the modelling accuracy and hence the performance. This could be achieved by a re-classification of the training data. Furthermore, the stable performance of models for a high number of rate classes adapted with a relatively small amount of training data indicates that the transformation the parameters undergo with speech rate variation might possibly be modelled as a continuous shift. Together with the better performance of the rate-adapted models that are based on general ones this suggests that speech rate variation can be modelled by means of a transformation based on a continuous measure.

7. References

[1] N. Mirghafori, E. Fosler, and N. Morgan, "Fast speakers in large vocabulary continuous speech recognition: Analysis and antidotes," in *Proc. European Conf. on Speech Communication and Technology*, 1995, pp. 491–494.

[2] J. P. Verhasselt and J. P. Martens, "A fast and reliable rate of speech detector," in *Proc. Int. Conf. on Spoken Language Processing*, 1996, pp. 2258–2261.

[3] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," in *Proc. European Conf. on Speech Communication and Technology*, 1997, pp. 2079–2082.

[4] F. Martínez, D. Tapias, and J. Álvarez, "Towards speech rate independence in large vocabulary continuous speech recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 1998, pp. 725–728.

[5] R. Faltlhauser, T. Pfau, and G. Ruske, "Creating Hidden Markov Models for fast speech by optimized clustering," in *Proc. European Conf. on Speech Communication and Technology*, 1999, pp. 407–410.

[6] T. Pfau and G. Ruske, "Creating Hidden Markov Models for fast speech," in *Proc. Int. Conf. on Spoken Language Processing*, 1998, pp. 205–208.

[7] A. Tuerk and S. Young, "Modelling speaking rate using a between frame distance metric," in *Proc. European Conf. on Speech Communication and Technology*, 1999, pp. 419–422.

[8] F. Martínez, D. Tapias, J. Álvarez, and P. Leon, "Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, 1997, pp. 469–472.

[9] M. Richardson, M. Hwang, A. Acero, and X. D. Huang, "Improvements on speech recognition for fast talkers," in *Proc. European Conf. on Speech Communication and Technology*, 1999, pp. 411–414.

[10] P. Delattre, "An acoustic and articulatory study of vowel reduction in four languages," *Int. Review of Applied Linguistics in Language Teaching*, vol. 7, pp. 295–325, 1969.

[11] B. Wrede, G. A. Fink, and G. Sagerer, "Influence of duration on static and dynamic properties of German vowels in spontaneous speech," in *Proc. Int. Conf. on Spoken Language Processing*, 2000, pp. 82–85.

[12] B. Lindblom, "Spectrographic study of vowel reduction," *J. Acoustical Soc. of America*, vol. 35, no. 11, pp. 1773–1781, 1963.

[13] M. Fourakis, "Tempo, stress, and vowel reduction in american english," *J. Acoustical Soc. of America*, vol. 90, no. 4, pp. 1816–1827, 1991.

[14] S. J. Moon and B. Lindblom, "Interaction between duration, context, and speaking style in english stressed vowels," *J. Acoustical Soc. of America*, vol. 96, no. 1, pp. 40–55, 1994.

[15] G. A. Fink, "Developing HMM-based recognizers with ESMERALDA," in *Lecture Notes in Artificial Intelligence*, V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, Eds., Berlin Heidelberg, 1999, vol. 1692, pp. 229–234, Springer.