On-line Speaking Rate Estimation Using a GMM/NN Approach

R. Faltlhauser and T. Pfau and G. Ruske
Institute for Human-Machine-Communication
Technical University of Munich (TUM), Germany
{faltlhauser,pfau,ruske}@ei.tum.de

Abstract

Gaussian Mixture Models (GMM) have become a popular tool in fields like speaker identification or verification. In these disciplines they are commonly used to model the general properties of a given speaker. In our paper we extend the application of GMMs to the assessment of speaking rate. The use of a single GMM layer allows the estimation of the speaking rate category. By adding a neural network (NN) layer, acting as a mapping function, a continuous measure for the speaking rate can be provided. Experiments with different types of feature vectors showed a correlation coefficient of up to 0.74 between the lexical phoneme rate and our estimation based on speech rate dependent spectral variation. It emerged that major information is contained in the derivative coefficients. As a spin-off, our approach can be used for simultaneous gender identification.

1 Introduction

Since speaking rate has tremendous influence on the recognition performance [1], it is necessary to do some kind of adaptation towards speaking rate or select an appropriate acoustic model set. A prerequisite therefore is to have a measure for the actual speaking rate.

For the measurement of the rate of speech (ROS) several criteria have been proposed. Most evident is the definition via the syllable or the phoneme rate, which are computed from a phonetic segmentation of the given utterance. Since a segmentation or a recognition output has to be available, they are not directly suitable for on-line application. Two types of approaches trying to overcome this problem can be distinguished: the first type aims at estimating the phonetic content e.g. vowels [3] or phoneme boundaries [4], whereas the second type comes up with features which are correlated with the speaking rate, e.g. energy envelope [2].

As e.g. Kuwabara [5] has shown, speaking rate has significant influence on the spectral characteristics of certain phonemes. In this paper we examined whether these spectral variations in spontaneous speech can be used to quantify speaking rate. Our approach is focused on the effects in feature space. Any speaking rate dependent artifacts on higher phonetic levels such as more frequent elisions or assimilations are not considered.

An excellent tool for modeling global properties without considering the underlying segmentation are so-called Gaussian Mixture Models (GMM). In the next chapter we present two ways how GMMs can be applied for this task, followed by the experimental results obtained.

2 ROS Estimation

2.1 Gaussian Mixture Models

Gaussian Mixtures Models (GMM) have proven to be a powerful tool for distinguishing acoustic sources with different general properties. This ability is commonly exploited in tasks like speaker verification or identification [6], where each speaker or group of speakers is modeled by a GMM. Their major advantage lies in the fact, that they do not rely on any phonetic segmentation of the speech signal. A fact that makes them ideal for on-line application. However, this advantage means at the same time, that they are usually not suitable for modeling temporal dependencies - but this disadvantage is of minor importance, if the focus lies on the representation of global spectral properties.

A Gaussian Mixture Model m is basically a weighted superposition of K Gaussian densities:

$$p(x|m) = \sum_{k=1}^{K} c_{km} N(x, \mu_{km}, \Sigma_{km},)$$

For each model $m = 1 \ldots M$ the mixture coefficients c_{km} have to obey the probabalistic constraint:

From: Tagungsband ITG-Fachtagung "Sprachkommunikation" 2000, ITG

$$\sum_{k=1}^{K} c_{km} = 1$$

During the recognition phase the scores (log. likelihoods) are accumulated for a sequence of feature vectors $X = \{x_1, \ldots, x_P\}$

$$S(X|m) = \sum_{j=1}^{P} log(p(x_j|m))$$

Finally the model is chosen yielding the highest likelihood score.

$$\hat{m} = \arg\max_{m} S(X|m)$$

2.2 Speaking rate category determination

In speaker identification tasks for example, a GMM is trained for each speaker. Similar to this setup, our approach uses one GMM per speaking rate category. For the task of rate category determination different speech rate categories are defined, ranging from very slow to very fast. In the case of 3 categories:

$$C \in \{slow, med, fast\}$$

or for 5 categories:

$$C \in \{xslow, slow, med, fast, xfast\}$$

In either case each category is represented by an individual GMM. During the recognition phase all GMMs are scored in parallel (Figure 1). By the time of evaluation the frame scores are accumulated. In order to provide continuously a measure for each incoming speech frame x_i the acoustic signal has to be windowed with window length N_W. Since only speech frames are necessary to determine the speaking rate all non-speech frames are discarded.

$$\bar{S}_i(X_i|m) = \frac{\sum_{j=i-N_W}^{i} log(p(x_j|m))\delta(x_j)}{\sum_{j=i-N_W}^{i} \delta(x_j)}$$

whereby

$$\delta(x_j) = \begin{cases} 1 & \text{if frame j is speech frame} \\ 0 & \text{else} \end{cases}$$

N_W can either be kept constant or - for a more robust estimation - it can be held dynamic such that:

$$\sum_{j=i-N_W}^{i} \delta(x_j) = N_{max}$$

which basically means that within an utterance always a constant number of speech frames is used for the estimation. Finally the category $C_{\hat{m}}$ belonging to the GMM with the highest accumulated score $\bar{S}_i(X_i|m)$ is selected as speech rate category hypothesis in frame i.

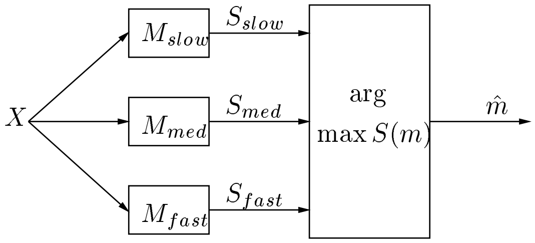

Figure 1: Maximum decision in case of 3 parallel category GMMs.

In order to account for the coarse acoustical differences caused by gender an approach with gender dependent GMMs is straightforward. Instead of one GMM per rate each speaking rate category is represented by two gender dependent (male/female) models. The recognition process basically stays the same: the category yielding the highest accumulated path score is chosen as a rate hypothesis. As a spin-off, this setup offers the advantage of simultaneous gender identification. In case gender is chosen as category, each gender is represented by 3 (respectively 5) GMMs. The definition of the category changes, now: $C = C_{gender}$, but the decoding step stays the same as with speaking rate.

2.3 Speaking rate estimation

A continuous measure might be useful for the adaptation of a recognition system towards speech rate. For this reason we replaced the decision logic by a deterministic mapping function f_{map} to calculate a continuous estimation for the speaking rate:

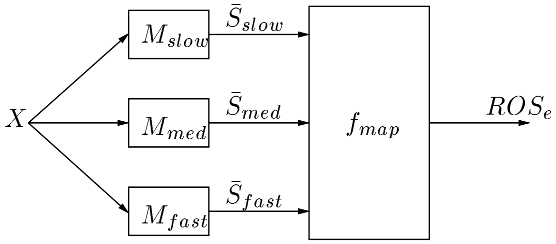

Figure 2: Continuous output for ROS_e realized by a NN layer f_{map}.

Hence, as inputs the accumulated and length normalized GMM output scores $\bar{S}$ are used. As mentioned beforehand we were examining an experimental setup with either 3 or 5 speaking rate GMMs. In case of 3 GMMs:

$$ROS_e = f_{map}(\bar{S}_{slow}, \bar{S}_{med}, \bar{S}_{fast})$$

The mapping function f_{map} is realized by means of a neural network consisting of at least 3 inputs, a possible hidden layer and an output layer which consists of a single neuron with linear activation function. For the input layer a hyperbolic activation function is chosen.

3 Experimental Results

3.1 Model training

The following experiments were carried out on the German Verbmobil database: about 11000 utterances of 600 male and female speakers. Each utterance was segmented into spurts, for which the phoneme rate $v = ROS_{Spurt}$ was calculated - leading to a total of about 33000 spurts. A spurt is basically a larger part of an utterance enclosed by non-speech segments. This segmentation seems necessary since the speech rate - especially in spontaneous speech - is often strongly varying over a complete utterance. For the evaluation we were using the eval96 test set consisting of 343 utterances (about 830 spurts).

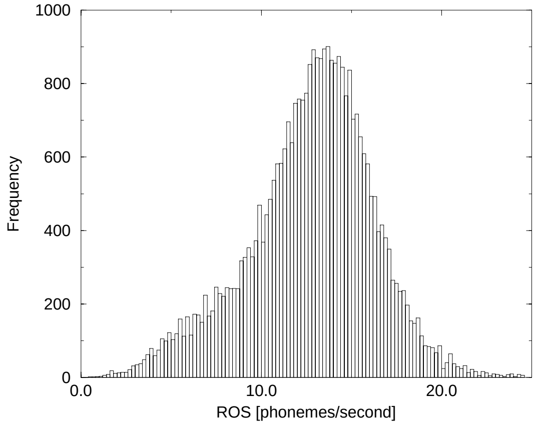

Figure 3: Histogram of phoneme rates (of spurts) in the training data.

In case of 3 rate categories $C = C(v)$ the category boundaries are given by $\mu \pm \Delta$:

$$C = \begin{cases} fast & \text{if } v > \mu + \Delta \\ slow & \text{if } v < \mu - \Delta \\ med & \text{else} \end{cases}$$

where $\mu = \mu_{ROS} = 12.765[phonemes/sec]$ and $\Delta = \sigma_{ROS} = 3.49[phonemes/sec]$.
Based on this classification each GMM is trained with the according data using Maximum-Likelihood (ML) estimation. In case of 5 different categories a second boundary at $2\sigma_{ROS}$ would leave too few training data for the "very fast/slow" GMMs, therefore we chose $\Delta = 2.0$.

$$C = \begin{cases} xfast & \text{if } v > \mu + 2\Delta \\ fast & \text{if } \mu + \Delta < v <= \mu + 2\Delta \\ slow & \text{if } \mu - 2\Delta <= v < \mu - \Delta \\ xslow & \text{if } v < \mu - 2\Delta \\ med & \text{else} \end{cases}$$

The inputs for the NN are taken from the outputs $\bar{S}_m$ of the GMM layer. For each spurt of the training data the resulting normalized output scores $\bar{S}_m$ are calculated and fed into the NN. As targets for the neural net the actual phoneme (or vowel) rates are provided. Training itself is performed using the backpropagation algorithm.

3.2 Speaking rate category determination

In order to see the influence of preprocessing we examined 3 different types of feature vectors:

- MFCC42
- MFCC12 and
- MUC66.

The first feature set consists of MFCCs with 42 dimensions: 12 MFCCs, total energy and zero crossing rate together with first and second derivatives. In MFCC12 only the MFCCs themselves are used. The last feature set is based on 20 linear, bark scaled features together with total energy, zero crossing rate and their derivatives.
Using GMMs with 16 mixture components the following classification results for 3 rate categories were achieved:

features:	corr. [%]	ρ
MFCC42	60.9	0.47
MFCC12	37.8	0.17
MUC66	35.5	0.43

Table 1: Classification results for 3 speaking rate categories.

Determining the rate category by a maximum decision yields a correlation coefficient $\rho \approx 0.45$ for the features including derivatives - the actual phoneme rate taken as reference. It can be seen from table 1 that major information contributing to the correlation with speaking rate originates from the delta coefficients.

features:	corr. [%]	ρ
MFCC42	35.3	0.58
MFCC12	23.4	0.23
MUC66	24.4	0.54

Table 2: Classification results for 5 speaking rate categories.

Table 2 shows basically the same result, although the overall correlation is higher, $\rho \approx 0.55$ for features including delta coefficients.

	estimated category		
given:	slow	med	fast
slow	75	8	2
med	196	373	69
fast	8	42	59

Table 3: Confusion matrix: actual against estimated speech rate category (MFCC42, 16 Gaussians)

	estimated category				
given:	xslow	slow	med	fast	xfast
xslow	58	8	5	1	1
slow	27	52	3	4	0
med	37	207	124	25	36
fast	3	26	59	22	53
xfast	2	6	23	12	38

Table 4: Confusion matrix: actual against estimated speech rate category (MFCC42, 16 Gaussians)

From the confusion matrices (tables 3, 4) we can see a high confusability between neighboring speech rate categories, e.g. medium and slow, but very few confusions between the opposing fast and slow categories. Considering the training data used for each category the high confusabilty between neighboring rate categories becomes understandable. As figure 3 shows, the data used for training are not separated by clear boundaries - in terms of speech rate - for the neighboring categories. Using a strict maximum decision therefore does not allow an exact classification according to the lexical phoneme rate, but it does of course reflect speech rate quite well.
Using a segmentation of the training as well as of the evaluation utterances it can be determined which phonemes contribute most in discriminating rate categories.

given:	training
slow	/m/, /n/, /e:/, /E:/, /2:/, /i:/, /a:/
fast	/U/, /aU/, /b/, /d/, /o:/, /z/
	evaluation
slow	/m/, /n/, /e:/, /E:/, /S/, /i:/
fast	/z/, /b/, /d/, /9/, /I/, /O/, /a/, /U/

Table 5: Phonemes with highest average score difference.

Table 5 shows for a given fast and slow category those phonemes which have on average the highest score difference if evaluated with the same category GMM and its opposing counterpart. Most of the phonemes are vowels.

3.3 Speaking rate estimation

Rate of speech (ROS) is commonly seen with respect to the lexical based definitions "phoneme rate" (ROS_{Ph}: phonemes per second) or "vowel rate" (ROS_V: vowels per second), which is strongly correlated with the lexical syllable rate. We focus in the following on the use of ROS_{Ph}. Basically the training of GMMs and the NN can be conducted as well with ROS_V, but the performance is slightly poorer as with ROS_{Ph}. In the following ROS depicts ROS_{Ph}.
The scatterplot in figure 4 shows the estimated phoneme rate (ROS_e) for each spurt of the evaluation data against the actual phoneme rate (ROS_a) calculated from the segmentation.

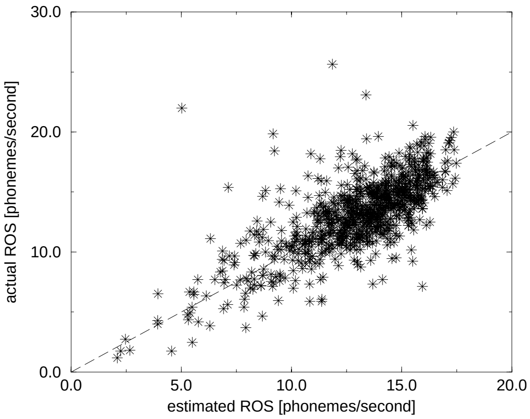

Figure 4: Scatter plot showing actual vs. estimated speaking rate.

Our ROS estimate reaches a correlation coefficient of $\rho = 0.74$ with ROS_a (table 6, 7) using 64 Gaussians per GMM. Some spurts have a fairly large deviation, as can be seen in Figure 4. This effect, especially in case of ROS_a, is primarily caused by the tradeoff between spurt length and the robust computation of ROS_a.

#Gaussians	1	16	64	256
MFCC42	0.624	0.745	0.745	0.738
MFCC12	0.080	0.308	0.305	0.340
MUC66	0.629	0.725	0.734	0,743

Table 6: Dependency of correlation coefficient ρ on GMM size for 3 categories.

#Gaussians	1	16	64	256
MFCC42	0.678	0.708	0.741	0.739
MFCC12	0.140	0.320	0.305	0.350
MUC66	0.604	0.698	0.726	0.731

Table 7: Dependency of correlation coefficient ρ on GMM size for 5 categories.

Figure 5 presents the distribution of the relative error $\epsilon = \frac{ROS_e - ROS_a}{ROS_a}$. It has zero mean and a standard deviation of 0.2.

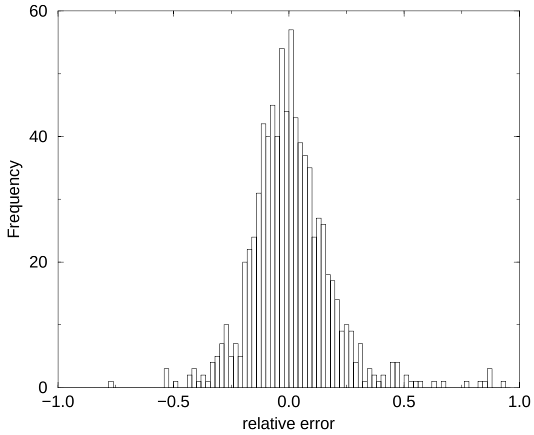

Figure 5: Histogram of relative error ϵ.

3.4 Gender identification

If using gender dependent GMMs our approach can simultaneously be applied for gender identification. In general, about 100 frames (1 second) of speech is sufficient to estimate the gender with high accuracy (table 8). Wrongly classified spurts are primarily caused by utterances shorter than 100 frames.

Gender recognition rate	
first 100 Frames of spurt	97.0%
entire spurt	99.1%

Table 8: Gender recognition rates.

4 Summary

Our results have shown, that at least slow and fast speech can clearly be distinguished already on the feature level. Furthermore the results confirm that major information about the speaking rate is contained in the delta coefficients. Since all GMMs are scored in parallel, the handling of our estimator offers various solutions: e.g. scoring during the speech spurts or a window based scoring scheme. However, the system relies on a robust speech-pause detector. Summarizing, our approach is fast, flexible and allows simultaneous gender identification as well.

5 Acknowledgements

This work was partially funded by the “Deutsche Forschungsgemeinschaft” (DFG).

References

[1] T. Pfau, G. Ruske, ”Creating Hidden Markov Models for Fast Speech”, Proc. ICSLP 98, paper 255, pp. 205-208

[2] N. Morgan, E. Fosler, N. Mirghafori, ”Speech Recognition using On-Line Estimation of Speaking Rate”, Proc. Eurospeech 97, pp. 2079-2082.

[3] T. Pfau, G. Ruske, ”Estimating the Speaking Rate using Vowel Detection”, Proc. ICASSP 98, pp. 945-948.

[4] J.P. Verhasselt, J.-P. Martens, ”A Fast and Reliable Rate of Speech Detector”, Proc. ICSLP 96, pp. 2258-2261.

[5] H. Kuwabara, ”Acoustic and Perceptual Properties of Phonemes in Continuous Speech as a Function of Speaking Rate”, Proc. Eurospeech 97, pp. 1003-1006.

[6] D. A. Reynolds, R. C. Rose, ”Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Trans. Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, January 1995.