ON-LINE SPEAKING RATE ESTIMATION USING GAUSSIAN MIXTURE MODELS

R. Faltlhauser, T. Pfau, G. Ruske

Inst. for Human-Machine-Communication
Technical University of Munich (TUM), Germany
Faltlhauser@ei.tum.de

ABSTRACT

Gaussian Mixture Models (GMM) are a widespread tool in applications like speaker identification or verification. In contrast to Hidden Markov Models (HMM) Gaussian Mixture Models are designed to model the general properties of an underlying acoustic source. In our paper we extend the application of GMMs to the assessment of speaking rate. Directly trained on the acoustic data, they can be either applied directly to estimate the speech rate category or - with the help of a mapping function - they can provide a continuous measure for the speaking rate. The mapping function can be realized by means of a Neural Net. First experiments showed a correlation coefficient of 0.66 between the lexical phoneme rate and our estimation based on speech rate dependent spectral variation. Moreover, our approach can be used simultaneously for high accuracy on-line gender detection.

1. INTRODUCTION

For the measurement of speaking rate (ROS) several criteria have been proposed. Most common are the definition via the syllable or the phoneme rate, which are computed from the phonetic transcription. Since the transcription or recognition output has to be available, they are not directly suitable for on-line application. Various other approaches have been proposed. Morgan et al. [1] introduced a measure based on the energy envelope, showing roughly a correlation of 0.5 with the phoneme rate. In [2] a feature called modified loudness, obtained directly from the speech signal is used to detect vowels in order to estimate the vowel rate, which is strongly correlated with the lexical syllable rate. Verhasselt et al. [3] used a Multi-Layer Perceptron to detect phoneme boundaries in order to estimate the phoneme rate.

Kuwabara [4] has shown that speaking rate has significant influence on the spectral characteristics of certain phonemes. In this study we wanted to see to what extent this effect is observable in spontaneous speech and how far it can be used to quantify speaking rate. Our aim was therefore to measure speaking rate based on speech rate dependent variations in feature space.

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, 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 [5], 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 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 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 composed of a superposition of K gaussian densities, whereby each density k is weighted with a mixture coefficient c_k:

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

The mixture coefficients have to obey for each model $m = 1 \ldots M$ the probabalistic constraint:

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

During the recognition phase the scores $log(p(x|m))$ are accumulated for the sequence $X = \{x_1, \ldots, x_P\}$

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

Finally the model is chosen, that yields the highest likelihood score.

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

In speaker identification tasks for example, a GMM is trained for each speaker. Our approach is very similar. In-

From: Proc. ICASSP 2000, IEEE

stead of speaker dependent we are using speech rate dependent GMMs.

2.2. Speaking rate category determination

For the task of rate category determination 3 speech rate categories are defined. Each category

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

is represented by an individual GMM (or respectively two GMMs: one male and one female). In order to train the models the training data is segmented into spurts, and for each spurt with known phonetic transcription the phoneme rate (or vowel rate) is calculated. A spurt is basically a larger portion 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.

All spurts with a speech rate exceeding $\mu_{ROS}+\sigma_{ROS}$ are marked as fast. In the same way all spurts falling below the symmetric threshold $\mu_{ROS} - \sigma_{ROS}$ are considered as slow. All remaining spurts are classified as medium. Based on this classification each GMM is trained with the according data using Maximum-Likelihood (ML) reestimation.

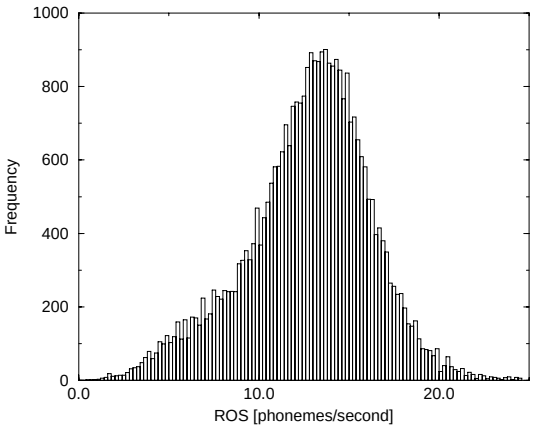

Figure 1: Histogram of phoneme rates (of spurts) in the training data.

$$C(ROS_{Spurt}) = \begin{cases} fast & \text{if } ROS_{Spurt} > \mu_{ROS} + \sigma_{ROS} \\ slow & \text{if } ROS_{Spurt} < \mu_{ROS} - \sigma_{ROS} \\ med & \text{else} \end{cases}$$

During the recognition phase all 3 GMMs are scored in parallel (Figure 2). By the time of evaluation the frame scores are accumulated. Finally the category $C_{\hat{m}}$ belonging to the GMM with the highest accumulated score is selected as hypothesis for the speech rate category.

As mentioned above an approach with six GMMs is also possible. Instead of one GMM per rate each speaking rate category is represented by two gender dependent (male/female) GMMs to separate the coarse differences in spectral representation caused by gender. Basically the

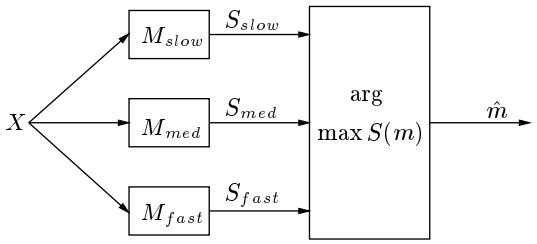

Figure 2: Maximum decision between 3 parallel GMMs.

recognition process stays the same: the category is recognized which yields the highest accumulated score. Remains to mention, that this approach offers another advantage: *gender detection can be done simultaneously.* In case gender is chosen as category, each gender is represented by 3 GMMs. Here only the definition of the category changes, now: $C = C_{gender} \in \{male, female\}$. The decoding step remains the same as with speaking rate. In the recognition phase the approach relies on a robust speech-pause detector, since only the speech frames are to be used.

2.3. Speaking rate estimation

For the adaptation of a recognition system to speech rate often a continuous measure for the speaking rate is needed instead of a discrete category. We therefore replaced the maximum decision with a deterministic mapping function f_{map} to calculate a continuous estimation for the speaking rate:

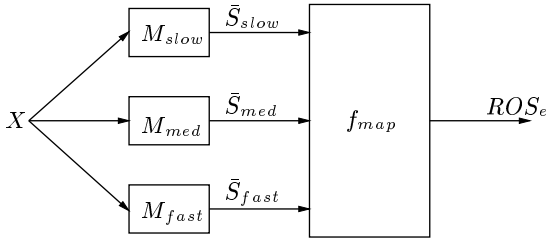

Figure 3: Continuous output for ROS_e realized by a Neural Network Layer f_{map}.

Hence, as inputs we are using the accumulated output scores normalized with the number of frames.

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

$$\bar{S}_m = \frac{1}{P} S(X|m)$$

The mapping function f_{map} is realized by means of a Neural Net consisting of 3 inputs, at least one hidden layer and an output layer, represented by a single Neuron with linear activation function (Fig. 4). The training inputs for the Neural Net 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 Neural Net. As targets for the Neural Net the actual phoneme (or vowel) rates are provided. Training itself is performed with the backpropagation algorithm.

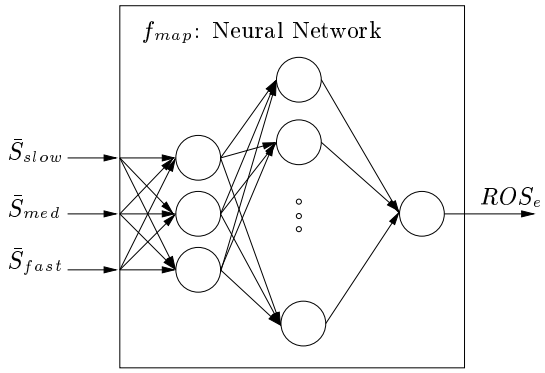

Figure 4: f_{map}: Neural Network with 3 inputs, 2 hidden layers and 1 output layer.

3. EXPERIMENTAL RESULTS

GMM training for the following experiments was performed on the German Verbmobil database: about 11000 utterances of nearly 600 male and female speakers. Each utterance was segmented into spurts, for which the phoneme (vowel) rate was calculated - leading to a total of about 33000 spurts. For the evaluation we were using the eval96 test set consisting of 343 utterances (about 830 spurts). Feature extraction is based on MFCCs with 42 dimensions: 12 MFCCs, total energy and zero crossing rate together with first and second derivatives. Using derivates causes, that temporal information is not completely disregarded.

3.1. Speaking rate category determination

Determining the rate category by a maximum decision yields about 64.5% correctly classified spurts - the actual phoneme rate taken as reference. This results in a correlation coefficient of about 0.47. Table 1 shows these results in detail.

	estimated category		
given:	slow	medium	fast
slow	79	71	3
medium	29	397	47
fast	4	141	61

Table 1: Confusion matrix: actual against estimated speech rate category.

From table 1 we see a very high confusability between neighboring speech rate categories, e.g. medium and slow. In contrast to this result there are very few confusions between the opposing categories fast and slow. The high confusabilty between neighboring rate categories becomes understandable considering the training data used for each category. As figure 1 shows, the data used for training are not separated by clear boundaries - in terms of speech rate - for the adjacent categories. Using a strict maximum decision therefore does not allow an exact classification according to the phoneme rate, but it does of course reflect speech rate quite well. Furthermore it gives rise to compute a continuous output.

3.2. 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. For the sake of simplicity we refer in the following chapter to both only as ROS, whereby ROS_a depicts the actual lexical based ROS and ROS_e refers to the estimate.

Phoneme Rate (ROS_{Ph}):

Figure 5 shows the estimated phoneme rate for each spurt of the evaluation data against the actual phoneme rate calculated from the segmentation.

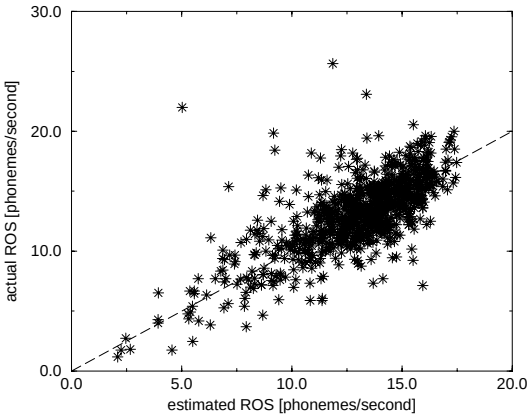

Figure 5: Scatter plot showing actual vs. estimated speaking rate.

Our ROS estimate reaches a correlation coefficient of $\rho = 0.66$ with the ROS_a, although some spurts have a fairly large deviation. This effect, especially in case of ROS_a, is primarily caused by the tradeoff between spurt length and the robust computation of ROS_a.

correlation coefficient ρ	
whole spurt	0.66
first 100 Frames	0.55

Table 2: Correlation coefficient ρ varies with evaluation window length.

ROS estimation beforehand was computed scoring a whole spurt. If the ROS estimate for a whole spurt was already drawn after the first 100 Frames (1 second) of the spurt the correlation coefficient still reaches $\rho = 0.55$. By using a running average this value could be improved.

Figure 6 presents a histogram over the relative error ϵ. It has zero mean and a standard deviation of 0.2.

$$\epsilon = \frac{ROS_e - ROS_a}{ROS_a}$$

A non-negligible effect is given by the model structure - i.e. number of gaussians - of the GMMs. Table 3 shows that the correlation between estimated and actual phoneme rate is dependent on the size of the GMMs. GMMs with

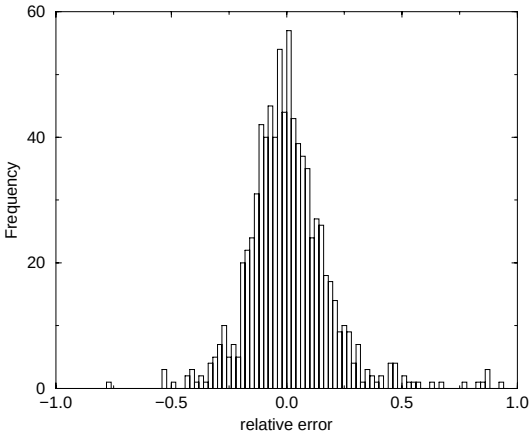

Figure 6: Histogram of relative error ϵ.

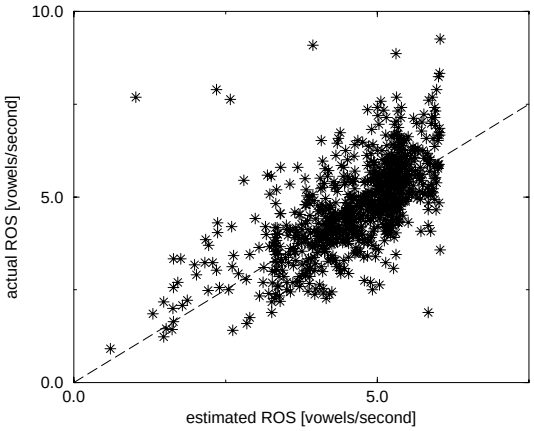

Figure 7: Scatterplot showing actual vs. estimated ROS.

32 gaussians achieve already a correlation of $\rho = 0.66$, but with the drawback of an increased variance. Using too many gaussians causes the correlation coefficient to decrease.

#gaussians per GMM	ρ	Std. dev σ_ϵ
128	0.648	0.206
64	0.665	0.193
32	0.660	0.220

Table 3: Correlation coefficient ρ varies with GMM model size.

Vowel Rate (ROS_V):

By replacing the targets for the Neural Network in the training phase, the system can be adapted to the estimation of the vowel rate. However, the results are slightly worse than with phoneme rate targets: the correlation coefficient between estimated and actual vowel rate drops to 0.59 (Fig. 7). Our estimate is based on the spectral variations of the frames of all phonemes, whereas the vowel rate is dependent on the phonemic contents. Since we evaluated our system on the spurt level, the calculation of the actual rates becomes more susceptible to the content of the phoneme string, especially for the vowel rate which depends on the syllable structure. This aggravates a robust computation of ROS_a here.

3.3. Gender estimation

As mentioned beforehand, our approach can - if using gender dependent GMMs - simultaneously be applied for gender detection. In general, about 100 frames (1 second) of speech is sufficient to estimate the gender with high accuracy (s. table 4). Wrongly classified spurts are primarily caused by utterances far shorter than 100 frames.

4. SUMMARY

First of all, our results have shown, that speaking rate has *significant* influence on the position of the vectors in feature space - proved by the fact, that slow and fast speech clearly can be distinguished already on the feature level. Since all GMMs are scored in parallel, the handling of the detection system offers various solutions: e.g. scoring during the speech spurts or a window based running average. However, the detection system relies on a robust speech-pause detector. Summarizing, our approach is fast, flexible and allows simultaneous gender detection as well. Further improvements through an optimized number of gaussians per GMM or a more sophisticated net structure seem possible.

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

Table 4: Gender recognition rates.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

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

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

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

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

[5] 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.