Failure Simulation for a Phoneme HMM Based Keyword Spotter

M. Holzapfel*, G. Ruske** and H.Höge*

** Siemens Corp. Otto Hahn Ring 6, D-81739 Munich, Germany*
*** Institute for Human Machine Communication, Technical University of Munich, Arcistraße 21, D-80290 Munich, Germany*
Martin.Holzapfel@mchp.siemens.de

Abstract

A basic problem in keyword spotting is the fact that the keywords itself cannot be completely different from background speech. Therefore, false alarms arise from those parts of the keyword which are also contained in the background. The paper describes the favourable application of a model trellis which enables to test individual phoneme sequences with respect to their influence on the underlying phoneme HMMs in a statistical way. It is shown, that the Viterbi path highly is affected by those partly fitting phoneme groups.

The probability of occurrance of these phoneme sequences is captured by a statistical "speech model" consisting of a Markov graph having an order up to 2. In this way sequences of 1, 2, or 3 phonemes are considered. By combining the model trellis and the statistical speech model, the probability of false alarms can be precalculated in advance, thus providing an useful measure for the suitability of the keyword under consideration. When the choice of keywords was optimized by this suitability measure in a practical application (spotting multicom 94.4 data) , the false alarm rate could be reduced by a factor of 3.5.

1. Introduction

In contrary to other architectures [2] [3] [4] [5] our previous keyword spotting system [1] used just a score threshold to indicate whether the keyword ends at a time or not. The scores have the meaning of a distance measure, that results in low scores in case of good matching. But low scores are not only caused by spoken keywords. So a more sophisticated analysis of the score caused by background speech is desirable. A main problem of keyword spotting is to separate spoken keywords from similar phoneme constellations appearing in background speech. These phoneme constellations may consist of phonemes or phoneme groups contained in the keyword. Most false alarms are caused by those parts of keywords. Normalizing the score by the length of the word hypothesis shows good results in detection, but increases this problem[1].

The choice of suitable keywords is a critical parameter for the performance of a word spotting system. The false alarm probability of a keyword is a major criterion for its suitability for word spotting.
In this paper a method for calculating this false alarm probability is presented.

2. Analysis of the course of the score

A model is introduced analysing the influence of any phoneme constellation in speech on the Viterbi path in the trellis and the score at the last state of the HMM.

The background score has been supposed to be an undetermined stochastic variable before. Now it can be shown, that the score is dominated by the influence of phonetic parts of the keyword.
That means that all cases have to be analyzed where phonemes of the the background coincide with phonemes of keywords.
This results in a hyperbolic decline of the score and a following relaxation. After the first phoneme of the keyword occurred in speech, the course of the score can be characterised by a sequence of those figures.

The length of the optimum path in the trellis also contains information about the word hypothesis. This information has not been used in former systems. In intervals up to 1000 ms, the length of the optimal path rises constantly, indicating those word hypotheses all starting at the same time. Their path in the trellis contains the same matching phoneme(s). Discontinuities in this length mark the change of the phoneme constellation dominating the trellis and in this way the score.

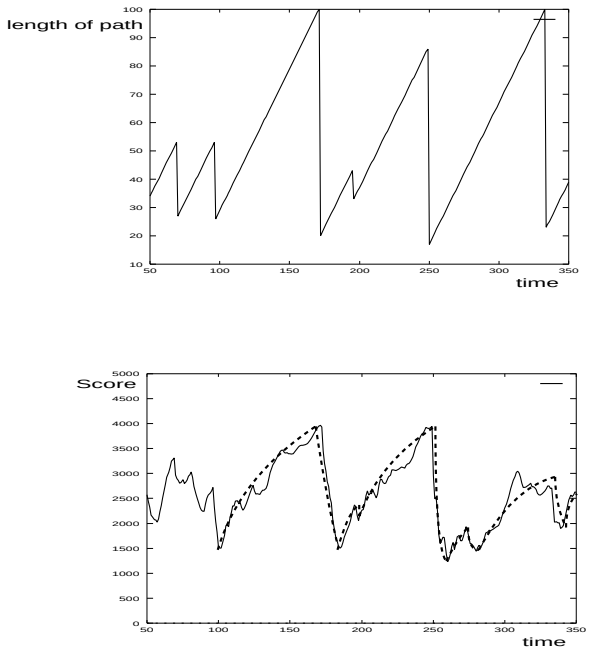

Figure 1: Score- and pathlength-history of a short speech sample

As an example score- and pathlength-histories of a short speech sample are shown in figure 1.

3. Modeling the influence of each single phoneme constellation

The course of score can be evaluated quantitatively in their statistic mean by simulation in a *model trellis*. This method is illustrated in figure 2. A hypothetic sequence of phonemes is distinguished in two cases: phonemes of the keyword and a generalised background. For this hypothetic speech signal artificial emission probabilities for each state of the HMM are derived for these two alternatives: part of the phoneme represented by this state spoken, or not spoken. These values, precalculated for each phoneme by statistical analysis of a large speech data base, are composed to an artificial model trellis with a virtual time axis, having the same dimensions as the trellis corresponding to a real HMM versus a real speech signal.

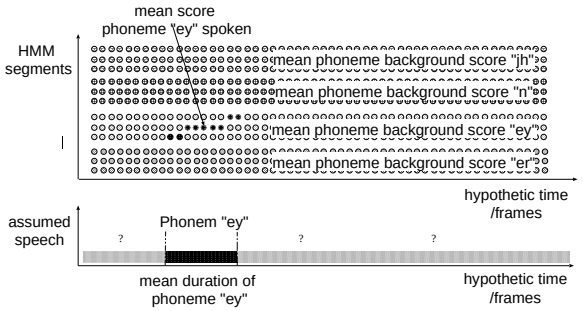

Figure 2: Model trellis for the keyword “er_ey_n_jh”

Using the Viterbi algorithm the mean score history can be computed. Figure 3 shows some characteristic paths and a mean decline and relaxation of the score, in simulation. The single phoneme “ey” is matching in this example in the keyword “arrange”.

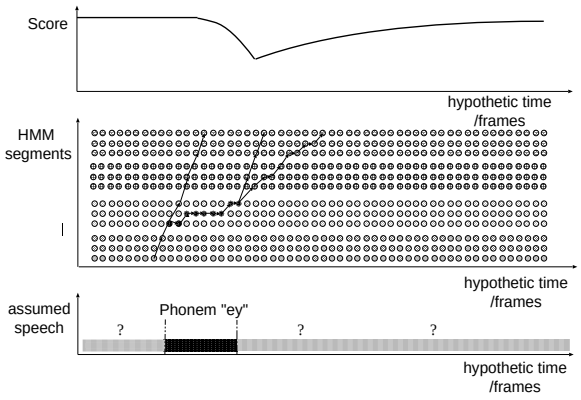

Figure 3: Score and model trellis for of phoneme “ey” matching

The minima of the score are points of special interest, as they are the candidates for detection as keyword. The distribution of probability of these values observed at the score minimum are estimated.

By backtracking, the path and its duration of staying in each phoneme model can be reconstructed.

The probability density functions of the phoneme-scores are assumed to be gaussian distributed. The pdf of the word score can be derived by length-weighted convolution. The basic principle is described in [1] but is extended to the really occurring phoneme sequences together with their statistics here.

As complexity of simulation has to be limited, all phonemes are assumed to occur by their mean length.

Analysis of the false alarms in real keyword spotting systems shows, that more than 97% of all false alarms are caused by one ore two contiguous groups of matching phonemes. In 95% of cases such a contiguous group is shorter than 4 phonemes (mainly syllables). It is sufficient to simulate constellations containing one or two groups consisting each of a maximum of three phonemes of the keyword. The distances between these groups to be simulated can be taken as a result of running modeling.
The number of cases to be simulated now, is of quadratic order to the number of phonemes in the keyword.

4. Probability of occurrence calculated by a speech model

Knowing the mean influence of each single phoneme constellation, it is desirable to estimate its probability of occurrence. A *speech model* is designed according to the restrictions of the model trellis, see figure 4.
Speech is considered to be a sequence of phonemes represented by a Markov graph up to second order. Each sequence of phonemes is represented by a path in the graph by a one-to-one correspondence. An extended waste class contains all phonemes and phoneme groups not being part of the keyword, plus nonarticulatoric events. Each path in the graph is assumed to start and end in this extended waste class.

To all other states the emission of one phoneme is assigned. Transition probabilities are represented by the arcs up to second order, if there are enough predecessors known, after having left the extended waste. According to modeling only chains of the length three, which are part of the keyword, the graph is very much reduced in its perplexity.

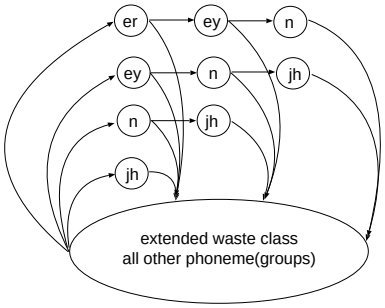

Figure 4: Speech model for calculations on parts of the keyword “er_ey_n_jh” e.g. “er_ey” and “jh”

Transition probabilities need to be calculated separately for each domain of speech by statistics on phonemes, phoneme tupels and -triples

5. Experimental results in application and conclusions

Using the model trellis and the speech model, a more precise pdf for the case „word not spoken“ can be calculated now:

$$Pdf(score|\overline{word}) =$$

$$\sum_{\text{all phonetic parts}} Pdf(score|\text{phon. parts}) \cdot P(\text{phon. part})$$

This definition of a *background pdf* allows a direct comparison of score minima of the spoken keyword and of background speech. These are the main opponents to be separated in classification [6].

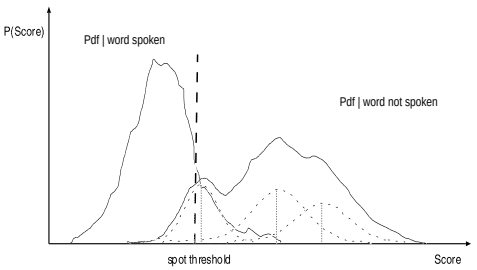

Figure 5: Pdfs for the cases “word spoken” and “word not spoken”

As a major quality criterion for the suitability of a keyword for word spotting, the false alarm probability can be estimated by this method. The correlation coefficient between precalculated and measured false alarms is 0.53 and allows only tendentious predictions. The correlation coefficient of the variance calculation [1] used at the model trellis is 0.57 . This can be regarded as theoretical upper bound in accuracy for estimation of the false alarm probability, which is nearly reached by the presented technique. In order to enhance the suitability measure, it is important to have provided more accurate estimates of this variance.

The calculation is able to distinguish keywords with quite similar " suitability by inspection", but a very different false alarmrate, as e.g. "evening" and "meeting". As choice of keywords is a critical parameter for accuracy of a word spotter, even quite raw estimation can be useful. In analogy to the law of big numbers, application of the keyword tester to a set of keywords is quite reliable.

For example using the five best keywords, out of the list of the twenty most frequent words in our test set (multicom 94.4), the false alarmrate is factor a 3.5 smaller than using the five worst keywords of this arbitrary list.

Based on this new quality of understanding the pathhistory in the trellis, new spot criteria can be derived utilizing both score and length of path.
A first, simple version is spotting just the score minima with respect to the influence of each phoneme constellation. Changes of the phoneme constellation dominating the path are indicated by discontinuities in the length of path.
This cirterion avoids double detection of one phonetic segment and reduces false alarm rate by a value up to 10%.

The mean trace of the path in the trellis can be predicted for each phoneme constellation by use of the model trellis. This detailed knowledge of path in trellis should enable more sophisticated spot criteria in future, as e.g. a weighted phoneme score criterion.

6. Acknowledgements

This work was partly founded by the German Ministry for Research and Technology (BMBF) in the framework of the "Verbmobil" Project. The responsibility for the contents lies with the authors.

7. References

[1] Junkawitsch J., Neubauer L., Hoege H., Ruske G., A new keyword spotting algorithm with pre-calculated optimal thresholds. ICSLP 96, Philadelphia, pp 2067-2070, 1996.

[2] Ohno S., Fujisaki H., Hirose K., A method for word spotting in continuous speech using both segmental and contextual likelihood scores. ICSLP, pp. 2199-2202, 1994.

[3] Hofstetter E. M., Rose R.C., Techniques for task independent word spotting in continuous speech messages. IEEE ICASSP, II, pp. 101-104, 1992.

[4] Chang, E. I., Lippmann P., Figure of merit training for detection and spotting. Advances in Neural Net Information Processing Systems 6, 1994, pp 1019 - 1026.

[5] Chigier B., Rejection and keyword spotting algorithms for a directory assistance city name recognition application. IEEE ICASSP 1992, II, pp 93 -96.

[6] Fukanaga K., Introduction to statistical pattern recognition, Academic Press, New York and London, 1972.