KEYWORD VERIFICATION CONSIDERING THE CORRELATION OF SUCCEEDING FEATURE VECTORS

Jochen Junkawitsch and Harald Höge

Siemens AG, Corporate Technology, D-81730 Munich, Germany
Email: Jochen.Junkawitsch@mchp.siemens.de

ABSTRACT

The assumption of statistically independent feature vectors within the HMM approach is a well known problem. The aim of this study is to explore a simple and feasible method, that takes the correlation of adjacent feature vectors into account. A so called correlated HMM, that estimates the emission probability of a state with respect to correlated feature vectors, is built by combining two separate knowledge sources. On the one side, a traditional HMM provides an emission probability under the condition of a certain state, whereas on the other side a linear predictor delivers an emission probability considering the previous feature vectors. The efficiency of this method is shown with the help of the German SpeechDat(M) database. The application of the correlated HMM within the verification procedure of a keyword spotter provided an improvement of the Figure-of-Merit from 87.1% to 88.6%.

1. INTRODUCTION

A variety of approaches have been investigated to overcome the assumption of statistically independent feature vectors. In [1] the correlation of features is explicitly considered and integrated into the HMM environment. This idea was often adopted and further developments and variants are reported, e. g. in [2] or [3]. Nevertheless, because of missing effectiveness or too high expense, the problem of correlated feature vectors still cannot be regarded as solved.

The aim of this study is to improve the keyword verification performance by taking the correlation of subsequent feature vectors into account. Therefore, a method is presented that enables a simple construction of a correlated HMM by combining probability estimates of two separate knowledge sources. This correlated HMM yields an additional score for each keyword hypothesis that is considered within the rejection procedure.

2. COMPOSING A CORRELATED HMM

The intention of this work is to present a simple and effective method for designing a correlated HMM. In order to overcome with the disadvantages of an integrated approach, the correlated HMM is build up with two separately handled knowledge sources.

The first knowledge source is a classical HMM without any changes, that covers the acoustic properties and provides an estimate for the emission probability $P_1(O_t)$ under the condition of a certain state by ignoring the correlation of adjacent feature vectors.

$$P_1(O_t) := P(O_t|q_t)$$

The second knowledge source regards the correlation of the feature vectors and yields an estimate of the feature probability $P_2(O_t)$ only using the previous features.

$$P_2(O_t) := P(O_t|O_{t-1}, O_{t-2}, \ldots)$$

These two probabilities are both estimates for the same but unknown "true" probability $P(O_t|q_t, O_{t-1}, O_{t-2}, \ldots)$ of a certain feature vector. In order to construct a correlated HMM, in a certain state a suitable combination (noted with the general operator "$\circ$") of these separate knowledge sources could yield a more accurate approximation

$$P'(O_t) = P_1(O_t) \circ P_2(O_t)$$

of the real distribution. The general problem of combining probability distributions is described in the literature, e. g. in [4]. The advantage of this method is obvious, because the representation of the acoustic-phonetic models have not to be modified. Neither sophisticated models have to be constructed in order to integrate the correlation of feature vectors, nor practical training algorithms that deal with the large number of parameters of such a model have to be invented. The usual HMM approach simply can be combined with a second knowledge source that considers the correlation.

3. ESTIMATING THE EMISSION PROBABILITY OF CORRELATED FEATURES

A linear predictor [5] is a suitable method to deal with correlated data sequences. It can be used to get an estimate for the probability of a certain feature vector O_t regarding the previous p feature vectors $\{O_{t-1}, \ldots, O_{t-p}\}$.

$$O_t = \hat{O}_t + \xi = \sum_{i=1}^{p} A_i O_{t-i} + \xi$$

The predictor coefficients A_i are matrices, $\hat{O}_t$ is an estimate for the real feature vector and ξ is a remaining prediction error vector.

As a further important simplification, a linear discriminant analysis is applied after the feature extraction, so that the single components of a feature vector are well decorrelated and no dependencies between the different dimensions of the feature vector must be considered by the predictor. As a consequence, the predictor may treat each component separately, and the predictor coefficients A_i become diagonal matrices.

The random variable O_t is the sum of the two random variables $\hat{O}_t$ and ξ. So the probability of O_t can be noted as a convolution of the corresponding probability density functions $P_{\hat{O}_t}$ and P_ξ. Since $\hat{O}_t$ is known from the predictor, its probability density function is a Dirac impulse with $P_{\hat{O}_t}(O_t) = \delta(O_t - \hat{O}_t)$. The remaining prediction error $\xi = O_t - \hat{O}_t$ can be considered to be a normal distributed random variable with mean $\mu_\xi = 0$, so that it can be expressed as:

$$P_\xi(\xi) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_\xi|}} e^{-\frac{1}{2}\xi^T \Sigma_\xi^{-1} \xi}$$

Due to the decorrelated vector components, the appropriate covariance matrix Σ_ξ is a diagonal matrix again. It can be determined concurrently with the parameter set A_i as the mean square discrepancies of the linear prediction process. So the resulting emission probability of a feature vector O_t can be expressed as follows:

$$\begin{aligned} P_2(O_t) &= P(O_t | O_{t-1}, \ldots, O_{t-p}) \\ &= P_{\hat{O}_t}(O_t) * P_\xi(O_t) \\ &= P_\xi(O_t - \hat{O}_t) = P_\xi(\xi) \end{aligned}$$

By the utilization of a linear predictor in this way, a emission probability of a feature vector can be calculated that considers the time correlation of subsequent feature vectors.

4. COMBINING PROBABILITY DENSITY FUNCTIONS

According to theory, e. g. [4], there are mainly two different approaches to combine probability distributions. The first can be described as a weighted sum of the particular density functions, so that the result is a multi-modal probability density function. The second possibility calculates the combined density as a weighted multiplication of the individual density functions. After re-normalization, uni-modal probability density functions are achieved.

Within this study, a little different approach is used. The combined probability density function is assumed to be a Gaussian, and its mean and variance are calculated by evaluating a weighted sum of the individual densities.

Some special properties of the underlying acoustic model have to be considered. The usual emission probability of a state of a HMM is modeled by a multi-modal density function, where all modes have a diagonal covariance matrix and all variances are equal, i. e. $\Sigma = I \cdot \sigma^2$. Moreover, the emission probability of the linear predictor has a diagonal covariance matrix, where all single variances are assumed to be equal, too, i. e. $\Sigma_\xi = I \cdot \sigma_\xi^2$.

With these assumptions, the weighted sum of the density functions of one mode (given by $\mathcal{N}(\mu, \Sigma)$) and the linear predictor (given by $\mathcal{N}(\hat{O}_t, \Sigma_\xi)$) can be computed as a two-modal density function.

$$(1-c) \cdot \mathcal{N}(\mu, \Sigma) + c \cdot \mathcal{N}(\hat{O}_t, \Sigma_\xi)$$

In order to obtain an uni-modal density, this weighted sum with $0 < c < 1$ is approximated by a single Gaussian by calculating the mean vector μ' and the global variance $\Sigma' = I \cdot \sigma'^2$ of this sum as

$$\mu' = (1-c) \cdot \mu + c \cdot \hat{O}_t$$

$$\sigma'^2 = (1-c)\sigma^2 + c\sigma_\xi^2 + (1-c)c(\mu - \hat{O}_t)^2$$

By this way, the emission probability of the correlated HMM is given by a Gaussian with mean μ' and variance Σ'. The combined probability is simply calculated by changing the mean and the variance of all modes of all states according to the above equations.

A further simplification is achieved by ignoring the modifications of the variances. The variance of each HMM prototype vector is assumed to be constant by setting $\Sigma' = \Sigma$. By this way, only a modification of the means is performed and the combined density function can be obtained as a Gaussian with a shifted mean vector $\mu' = (1-c) \cdot \mu + c \cdot \hat{O}_t$.

5. KEYWORD VERIFICATION

The goal of this work is to examine the usefulness of the correlated HMM for keyword verification and rejection. For

this purpose the proposed methods are tested using a two pass keyword spotting system.

At the first pass keyword hypotheses are generated using a modified Viterbi-algorithm, that is described more detailed in [6]. This method is based on local confidence scores instead of acoustic probabilities and works by optimizing a length-normalized confidence measure for each keyword separately. With t_1 and t_2 as keyword boundaries, this confidence measure is defined by the following likelihood ratio:

$$x_1 = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} -\log\left(\frac{P(O_t|q_t)}{P(O_t|\overline{q_t})}\right)$$

The denominator of this equation may be regarded as the emission probability of an only assumed anti-state $\overline{q_t}$. Its score (in the log-domain) can be approximated very well by averaging the n best state scores according to

$$-\log P(O_t|\overline{q_t}) = \tfrac{1}{n}\textstyle\sum_n (-\log P(O_t|q_n)).$$

As a result, this hypotheses generation process delivers keyword hypotheses, that are rated by the above confidence measure x_1. When working without additional rejection criteria, this value is the sole base for deciding between keyword acceptance and rejection by a comparison with a certain threshold.

The intention of the second verification pass is to ascertain a second rejection criterion, so that a more advanced keyword verification can be done. Similar to the above confidence measure, the second rejection criterion is defined as a length-normalized likelihood ratio, where the classical probability $P(O_t|q_t)$ is replaced by the probability P' of the correlated HMM.

$$x_2 = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} -\log\left(\frac{P'(O_t|q_t, O_{t-1}, \ldots, O_{t-p})}{P(O_t|\overline{q_t})}\right)$$

As an simple and feasible approach, the correlated HMM is produced by only shifting the means of all densities towards the vector that is conceived from the linear predictor according to the above idea. The variances of the densities are ignored and remain unchanged.

By this way, every keyword hypothesis can be re-rated using a linear combination of both rejection criteria as a new score.

$$score = w_1 x_1 + w_2 x_2$$

6. EXPERIMENTS AND RESULTS

The German SpeechDat(M)[1] database, that was recorded via the public telephone network, is used for testing the efficiency of the proposed correlated HMM for keyword verification. A total number of 22136 utterances from 667 speakers were taken to train a general context dependent HMM-set. The goal was to detect keywords within the so-called application phrases, that are a specific part of the database. A subset of 428 application phrases from different 167 speakers is used for testing purposes.

[1] For information about SpeechDat see the following URL's:
http://www.phonetik.uni-muenchen.de/SpeechDat.html
http://www.icp.grenet.fr/ELRA/home.html

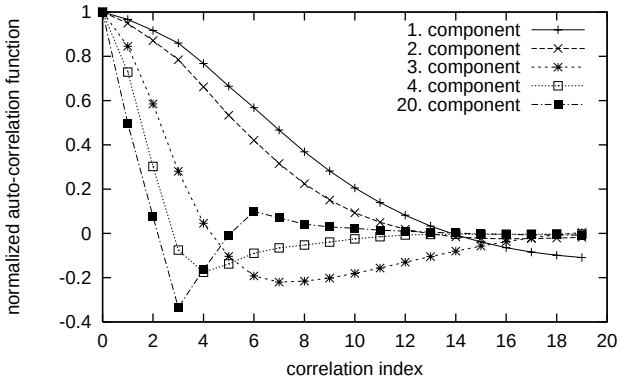

Figure 1: Normalized auto-correlation functions of selected feature vector components

Feature extraction is performed at a sampling rate of 8 kHz by calculating a total number of 24 mel-filtered cepstral coefficients. In order to compensate different channel transfer characteristics, a maximum likelihood based cepstral mean removal technique is applied to this 24 dimensional vector. Adding 12 first and 12 second order derivatives and including an energy component with its both derivatives, a 51 dimensional vector is composed. By combining two subsequent vectors at each time frame, a 102 dimensional super-vector is obtained, which is transformed using linear discriminant analysis. Finally, the resulting feature vector is determined by selecting the first 24 components out of the transformed and ordered super-vector.

In a first experiment the dependencies of adjacent feature vectors are explored by calculating their auto-correlation function, that are shown in figure 1. Although feature extraction is performed by adding first and second order derivatives with a subsequent linear discriminant analysis using super-vectors, it turned out that the feature vectors are still correlated in time. The most important components, indicated by high eigenvalues, show a higher degree of correlation even over a longer time period than the less important components with lower eigenvalues. On the other hand, the single components of a particular feature vector are decorrelated well. In order to regard this correlation of adjacent feature vectors and to design a correlated HMM, a linear predictor with $p = 4$ was determined.

In a second experiment the influence of the weighting factor c is investigated by testing distinct values of c. The

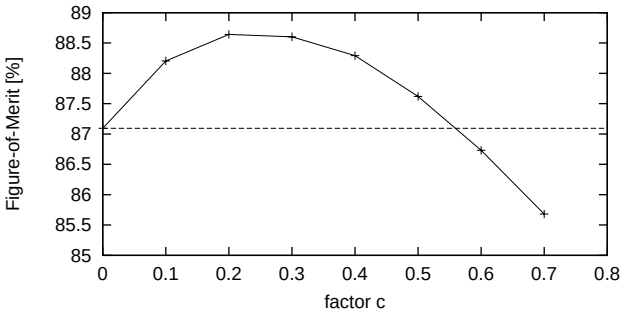

Figure 2: Figure-of-Merit corresponding to weighting factor c

weighting factors for re-scoring the keyword hypotheses are chosen as $w_1 = w_2 = 0.5$, i. e. the importance of both particular scores from the classical and the correlated HMM are balanced.

Figure 2 shows the Figure-of-Merit and the corresponding weighting factors c. The factor $c = 0$ stands for no shifting of Gaussian means and so the correlated HMM is identical to the classical one. This values should serve as a benchmark of the reference system. Best results are achieved by choosing $c = 0.2$, where a FOM improvement of 1.5% is reached.

In figure 3 the receiver-operating-characteristic (ROC) for a weighting factor $c = 0.2$ is shown and compared to the reference system. The most profits are yielded in the range of low false alarm rates where the differences of the detection rates have a maximum. Obviously the detection rates must converge for high false alarm rates, because the correlated HMM is only used for verification and rejection and does not improve the performance of the preceding hypotheses generation process.

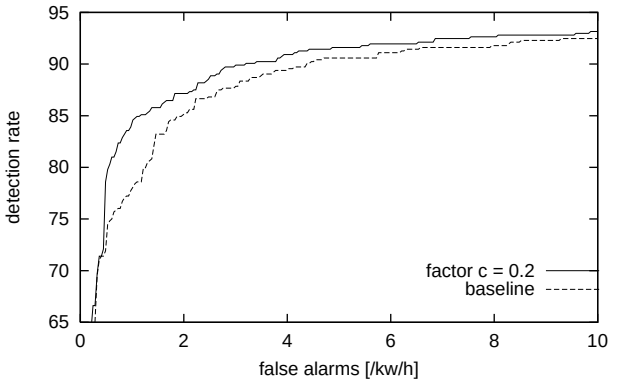

Figure 3: Receiver-Operating-Characteristic (ROC) for a weighting factor $c = 0.2$, compared with the baseline system

7. SUMMARY AND DISCUSSION

The objective of this paper is to investigate the efficiency of a correlated HMM for verification and rejection purposes. Therefore, a correlated HMM is defined by combining probability density functions from separate knowledge sources. The one is the acoustic-phonetic side, which is given by a traditional HMM. The other knowledge source considers the correlation of feature vectors and estimates the emission probability with respect to the previous feature vectors. This task is done by a linear predictor. Moreover, a method for combining these two estimates is introduced, that provides a simple and feasible calculation of the correlated HMM by shifting the means of all HMM prototypes.

Experiments with the German SpeechDat(M) database yield an 1.5% increase of the Figure-of-Merit, when the correlated HMM is applied for getting an additional rejection criterion, that is used for re-scoring the keyword hypotheses. This improvement is achieved, although some techniques are involved, that could have negative effects on correlated approaches, e. g. the application of delta components, a channel compensation procedure, and a linear discriminant analysis using super-vectors.

Moreover, the combination of the two probability density functions without respect to any variances is a simplification, where a more detailed approach further could enhance the success of the proposed methods.

8. REFERENCES

[1] C. J. Wellekens. Explicit time correlation in Hidden Markov Models for speech recognition. In *Proc. ICASSP*, volume 1, pages 384–386, 1987.

[2] P. Kenny, M. Lennig, and P. Mermelstein. A linear predictive HMM for vector-valued observations with applications to speech recognition. *IEEE Transactions on acoustics, speech and signal processing*, 38(2):220–225, February 1990.

[3] N. S. Kim and C. K. Un. Frame-correlated Hidden Markov Model based on extended logarithmic pool. *IEEE Transactions on speech and audio processing*, 5(2):149–160, March 1997.

[4] C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–148, 1986.

[5] J. D. Markel and A. H. Gray, Jr. *Linear Prediction of Speech*. Springer Verlag, Berlin, Heidelberg, New York, second edition, 1980.

[6] J. Junkawitsch, G. Ruske, and H. Höge. Efficient methods for detecting keywords in continuous speech. In *Proc. EUROSPEECH*, volume 1, pages 259–262, 1997.