

## A NEW APPROACH TO ACOUSTIC-PHONETIC DECODING BY MEANS OF HIDDEN-MARKOV-MODELLING OF SPEECH

B. Keck\*

### ABSTRACT

A connected speech recognition method based on stochastic modelling of speech is presented. The speech signal is segmented and labelled by a Viterbi-algorithm that finds the optimal path in an acoustic-phonetic Hidden-Markov-Model (HMM) of the language. This HMM is an integration of speech subunit models and a language model on the subunit level.

### INTRODUCTION

The problem of connected speech recognition is to find the best string of speech units by optimally matching the utterance to every possible concatenation of speech unit reference patterns or speech unit production models. The speech units might be words (connected word recognition); for the recognition of continuous speech with a large vocabulary it is useful to extract smaller units (phoneme-like units) from the speech signal ('phonetic engine'). In this case the the estimated subunit string is the input of the other (high-level) components of an integrated speech understanding or speech recognition system.

If  $Y = \{y_1, y_2, \dots, y_T\}$  is a sequence of acoustic vectors and  $W = \{w_1, w_2, \dots, w_K\}$  is a possible subunit string, the problem of recognition is to find the most probable subunit string  $\hat{W}$  that maximizes the conditional probability  $P(W|Y)$ .

Applying Bayes' rule  $\hat{W}$  must be chosen so that the conditional probability

$$P(\hat{W}|Y) = \max_w [P(W|Y)] = \max_w [P(Y|W) * P(W) / P(Y)] \quad (1)$$

Since  $P(Y)$  does not depend on  $W$ , we have to find the string that maximizes the product  $P(Y|W) * P(W)$ .

### PHONEMIC LANGUAGE MODEL

To solve the maximisation problem (1) it is necessary to know the a-priori probabilities  $P(W)$  of any subunit string from a language model on the subunit level. It seems to be a good idea to approximate  $P(W)$  as the product of the probabilities of all subunit digrams of the string  $W$

$$P(W) = \prod_{k=1}^K P(w_k | w_{k-1}). \quad (2)$$

We used phoneme-like subunits and we estimated the a-priori-probabilities of the subunit digrams by using programs that automatically transform a grapheme string to a subunit string ('phoneme string'). A large German text (size about 160000 words) was transformed in order to calculate the relative frequencies of the subunit digrams that were taken as estimates of the respective probabilities.

\*Fraunhofer-Institute IAO, Holzgartenstr.17, D-7000 Stuttgart 1, Germany

## SUBUNIT HIDDEN-MARKOV-MODELS

Hidden Markov modelling is an adequate method to describe the production of speech units. The parameters of a HMM of a speech unit are the transition probabilities  $a_{ij}$  and the probabilistic density functions  $b_i$ . The transition probabilities  $a_{ij}$  describe the transitions between states  $i$  and  $j$ ; the density functions  $b_i$  describe the acoustic vectors that are emitted when state  $i$  is visited.

To represent the phoneme-like subunits we used HMMs with 1,2,3 or 4 states depending on the acoustic properties of the subunit. One state seems enough to model very short units like many consonants, three states are adequate to model most vocals (with no loop in the first and in the last state representing the transitions from and to the neighbouring phonemes) and four states for the more transient diphthongs. We used Gaussian emission density function  $b_i$  with complete covariance matrices; the acoustic vectors (12 cepstral coefficients plus normalized signal energy) were calculated every 10 ms.

The training of the subunit-HMMs can be performed simultaneously for the models of all speech units by using known utterances of connected speech (embedded training). Such a procedure rests on the iterative forward-backward-algorithm. If the initial parameter estimates are good enough it is possible to apply the forward-backward-algorithm to larger utterances like sentences to improve the subunit-HMM parameters (fig. 1) because left-to-right subunit-HMMs can easily be concatenated according to the known subunit string.

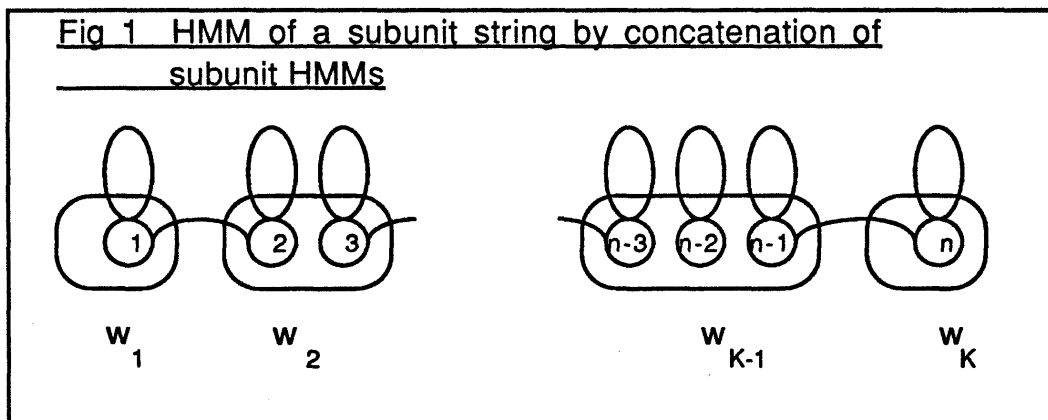
Using a HMM of a subunit string  $W$  (see fig.1) makes it easy to calculate the conditional probability  $P(Y|W)$  to solve the decoding equation (1). This probability can be calculated with the forward-backward-algorithm or approximatively with a Viterbi-algorithm which gives the probability of the best path in the HMM of the subunit string. With the initial state distribution vector  $[c_0(1), c_0(2), \dots, c_0(n)]^T = [1, 0, \dots, 0]^T$  the Viterbi-algorithm perform the following recursion for all times  $1 \leq t \leq T$  and all states  $1 \leq j \leq n$

$$c_t(j) = \max_i [c_{t-1}(i) * a_{ij}] + b_j(y_t) \quad . \quad (3)$$

The conditional probability  $P(Y|W)$  is approximatively

$$P(Y|W) \sim \hat{P}_1 = \max_i [c_T(i)] \quad , \quad (4)$$

which is the probability of the best state sequence  $S = s_0, s_1, \dots, s_T$ . This best state sequence or best path can be recovered by a backtracking procedure.



## FIRST DECODING ALGORITHM

With equations (2),(3) and (4) the decoding equation (1) could be solved; but in practice it will not be possible to calculate the product  $P(Y|W)*P(W)$  or  $P_1*P(W)$  for all possible subunit strings. Now let us assume to have solved equation (1) by applying (2),(3) and (4) to get the best subunit string  $W = \hat{w}_1, \hat{w}_2, \dots, \hat{w}_K$ . The probability of the best path in the HMM of  $W$  is

$$\hat{P}_1 = a_{s_0s_1} * b_{s_1}(y_1) * a_{s_1s_2} * b_{s_2}(y_2) * \dots * a_{s_{T-1}s_T} * b_{s_T}(y_T). \quad (5)$$

From (5) and (2) the probability of the product to be maximized is

$$\hat{P}_1 * P(\hat{W}) = \left( \prod_{t=1}^T a_{s_{t-1}s_t} \right) * \left( \prod_{t=1}^T b_{s_t}(y_t) \right) * \left( \prod_{k=2}^K P(\hat{w}_k | \hat{w}_{k-1}) \right). \quad (6)$$

## OPTIMIZED DECODING ALGORITHM

Now let us show that the same result can be achieved in a more straightforward manner. Let us build a HMM as an integration of all the subunit-HMMs and the language model on the subunit level as demonstrated in fig.2. The states of this HMM are the  $N$  states of all the subunit HMMs. The emission density functions of this integrated HMM are the density functions of the constituting subunit-HMMs. The transition probabilities  $\hat{a}_{ij}$  (loop in state  $i$ ) are the  $a_{ij}$  from the subunit-HMMs

$$\hat{a}_{ij} = a_{ij} \quad ; \quad (7a)$$

the transition probabilities between different states  $i$  and  $j$ ,  $i \neq j$ ,  $1 \leq i \leq N$ ,  $1 \leq j \leq N$  are

$$\hat{a}_{ij} = 1 - a_{ii} \quad \text{if } i \text{ and } j \text{ are different states of the same subunit-HMM} \quad (7b),$$

$$\hat{a}_{ij} = (1 - a_{ii}) * P(w_j | w_i) \quad \text{if } i \text{ is the last state of unit } w_i \text{ and } j \text{ the first state of } w_j \quad (7c),$$

$$\hat{a}_{ij} = 0 \quad \text{otherwise.}$$

We can use the Viterbi-algorithm to get the best path in this integrated HMM. The probability of this best path  $S = s_0, s_1, \dots, s_T$  is with (3) and (4)

$$\hat{P}_2 = \left( \prod_{t=1}^T \hat{a}_{s_{t-1}s_t} \right) * \left( \prod_{t=1}^T b_{s_t}(y_t) \right). \quad (8)$$

Let the subunit sequence corresponding to the optimal state sequence  $S$  be  $V = \hat{v}_1, \hat{v}_2, \dots, \hat{v}_K$ , then

$$\hat{P}_2 = \left( \prod_{t=1}^T a_{s_{t-1}s_t} \right) * \left( \prod_{k=2}^K P(\hat{v}_k | \hat{v}_{k-1}) \right) * \left( \prod_{t=1}^T b_{s_t}(y_t) \right). \quad (9)$$

Since the Viterbi-algorithm is proved to find the maximal probability  $\hat{P}_2$  must be identical to  $\hat{P}_1$  from (6); therefore we can conclude that  $\hat{v} = \hat{w}$ . We have

solved equation (1) by simply using a Viterbi-algorithm and an integrated HMM (fig.2).

## RESULTS

The decoding algorithm was tested in a speaker-dependent phoneme recognition experiment (39 German phoneme-like units). The rate of confusion, deletion and insertion errors was about 16% each. Better results might be achieved using improved subunit models and by improving state duration modelling. Since the Viterbi-algorithm can be implemented using additions and max-operations only, the presented decoding algorithm is suitable for real-time applications.

## REFERENCES

1. S E Levinson, L R Rabiner, M M Sondhi, The Bell Systems Technical Journal, Vol 62, No 4, pp 1036-73, April 1983.
2. A J Viterbi, IEEE Trans. Information Theory, IT-13, pp 260-9, april 1967.

