THE USE OF META-HMM IN MULTISTREAM HMM TRAINING FOR AUTOMATIC SPEECH RECOGNITION

C.J.Wellekens, J.Kangasharju, C.Milesi
Institut Eurécom - Sophia Antipolis - France*
(Christian.Wellekens@eurecom.fr)

ABSTRACT

Among the different attempts to improve recognition scores and robustness to noise, the recognition of parallel streams of data each one representing partial information on the test signal and the fusion of the decisions have received a great deal of interest. The problem of training such models taking recombination constraints at the level of speech-subunits has not yet been rigorously addressed. This paper shows how equivalence with an extended meta-HMM solves the problem and how reestimation formulas have to be applied to guarantee equivalence between the multistream model and the meta-HMM.

1. INTRODUCTION

Since speech is a non-stationary signal, it is analyzed over small time windows where local stationarity is assumed. A typical width for these windows is 10 ms and different kinds of analysis can be conducted on them. The first standard approach is a harmonic analysis that could be obtained using either filter banks or FFT transforms. After elimination of the fundamental period (pitch), a smoothed spectrum may be decimated resulting in a feature vector which size of is several tenths of entries corresponding to the power in frequency bands. Many other representations of the speech signal are possible leading all to vectors associated with a 10msec time slot such as LPC prediction coefficients or cepstral coefficients. Speech units (phonemes, diphones, syllables, words) are represented as HMM and to recognize an utterance, it is matched versus the "best" sequence of HMM. This matching is achieved by associating each feature vector of this utterance to one state of the HMM according to the best path. A probability density function gives a measure of the closeness of a vector and a state and is referred to as emission probability. Sentence recognition relies on the global probability which is the product of all emission probabilites and transition probabilities met along the best path in the HMM.

Obviously, such a process imposes that all entries of a current feature vector are associated with the same state. A drawback is that any perturbation of an entry of a vector increases the distance of this vector from that state even if all other entries match perfectly: this is particularly well observed if the speech signal is represented with smoothed spectrum vectors and distorted by a narrow band noise which affects only one entry [1-3].

The idea came to split the feature vectors and to separately process all entries or group of entries with different HMM's: in this case, the denomination subband instead of multistream recognizers is used. Allocating different weights to the streams may enhance scores by weighting down noisy streams.Multistream terminology generalizes the process to the cases where feature vectors contain other informations than spectral information or even informations consistent with the utterances like visual information on the lip contours. However, this paper intends neither to compare different techniques for improvement of ASR robustness nor to describe applications of visual speech but to address a specific training problem of multistream models.

Imposing synchronism of all entries of feature vectors on states may be questioned. Multistream HMM allows a certain degree of asynchronism inside subunits. Complete asynchronism at the level of the whole sentence would prevent recognition in words which is based on the segmentation by the optimal path: for that reason asynchronism is restricted inside subunits.

As a consequence, in the utterance of a sentence, there exist some points where the parallel flows of the multistream HMM must recombine: at these points, we may ask the question: how must we recombine the scores of the streams? A second question is: how can we train a multistream HMM while forcing recombination at some points (undefined before the optimal alignment of the utterance over the states) as it is necessary at the end of subunits (at least at the end of words but also maybe at the end of syllables or phonemes).

*Institut Eurécom is supported by Ascom, Cegetel, France Telecom, Hitachi, IBM, Motorola, Swisscom,Texas Instruments, Thomson CSF

These two questions are addressed in the paper. In particular, it will be shown that making the product of the probabilities corresponding to the streams yields a model which fits with the simple stream model and leads to a complete equivalence with a meta HMM model. The meta-model which will be trained must meet strict contraints on its transition probabilities in order to have an equivalent multistream model.

Other techniques have been proposed to train multistream in particular [4] which uses a method derived from HMM decomposition. In section 2, a 2-stream-HMM model is described each channel being a 2-state Bakis model.Generalization to multistream with more sophisticated channel models is straightforward. Equivalences between emission probabilities with those of a meta-HMM enforcing recombination are derived. Section 3 addresses the equivalence of transition probabilities and shows that strict restrictions apply on the transitions of a meta-HMM in order to be equivalent to a multistream model. In section 4, a procedure for training meta-HMM which garantees the equivalence is proposed. In section 5, report is made on tests and preliminary conclusions are drawn.

2. THE META-HMM

Let us consider the case of a 2-stream HMM describing a phoneme and where each channel contains 2 states (fig 1) denoted respectively s_1 and s_2 for the upper stream and s_3 and s_4 for the lower stream. Transition loop probabilities on states s_i are denoted p_i while the state output probabilities are $q_i \overset{\text{def}}{=} 1 - p_i$. Since training as well as recognition implies that recombination occurs at the end of each subunit (here we consider phonemes) in the optimal path, it is necessary to create a meta-HMM described as follows. States of the meta-HMM are referred to as meta-states (s_{ij}). Visiting state s_{11} means that the upper and lower parts of a feature vector are emitted respectively on s_1 and s_3 and similarly for the other states according to table I.

	upper vector ξ	lower vector η
s_{11}	s_1	s_3
s_{12}	s_1	s_4
s_{21}	s_2	s_3
s_{22}	s_2	s_4

Table I: Correspondance between meta-states and multistream model states.

Self-loops are allowed on all states of the meta-HMM and transitions exist from s_{11} to all states (p_{11-ij}) and from s_{12} (p_{12-22}) and s_{21} (p_{21-22}) to s_{22}. There is only one output transition p_{22-F} from state s_{22} which enforces recombination.

Let us try now to find the relation between the parameters of the multi-stream model and the meta-HMM. The dimension of vectors in both streams are respectively d_u and d_l while the dimension of analyzed feature vectors is $d = d_u + d_l$. For the sake of clarity, we assume that the probability density functions associated with each state s_i are gaussian. The dimensions of the mean vectors μ_i and covariance matrices Σ_i associated with states s_i depend of course on the stream which they belong to (d_u and d_l).

In the alignment of a sequence of N vectors $x_k^T = (\xi_k^T, \eta_k^T)$ on the multistream model, we may consider that n_i vectors are associated with state s_i and recombination implies

$$n_1 + n_2 = n_3 + n_4 = N > 2, \qquad (1)$$

N may be equal to 2 only in case of full synchronism. The total probability of the upper stream is then

$$\prod_{k=1}^{n_1} p(\xi_k|s_1) \prod_{k=n_1+1}^{N} p(\xi_k|s_2) p_1^{(n_1-1)} p_2^{(n_2-1)} q_1 q_2 \qquad (2)$$

and for the lower stream

$$\prod_{k=1}^{n_3} p(\eta_k|s_3) \prod_{k=n_3+1}^{N} p(\eta_k|s_4) p_3^{(n_3-1)} p_4^{(n_4-1)} q_3 q_4. \qquad (3)$$

The global probability of the multistream models takes different expressions according to the kind of path in the model. In case $n_1 > n_3$, only states s_{11}, s_{22}, s_{12} of the meta-HMM are visited, there is a certain amount of asynchronism and the cumulated probabilities are respectively for the emission probabilities and transition probabilities

$$\prod_{i=1}^{n_3} p(x_i|s_{11}) \prod_{i=n_3+1}^{n_1} p(x_i|s_{12}) \prod_{i=n_1+1}^{N} p(x_i|s_{22})$$

$$p_{11-11}^{(n_3-1)} p_{11-12} p_{12-12}^{(n_1-n_3-1)} p_{12-22} p_{22-22}^{(n_2-1)} p_{22-F},$$

while if $n_1 < n_3$, only states s_{11}, s_{22}, s_{21} are visited, the paths in the streams are still asynchronous and the cumulated probabilities are

$$\prod_{i=1}^{n_1} p(x_i|s_{11}) \prod_{i=n_1+1}^{n_3} p(x_i|s_{21}) \prod_{i=n_3+1}^{N} p(x_i|s_{22})$$

$$p_{11-11}^{(n_1-1)} p_{11-21} p_{21-21}^{(n_3-n_1-1)} p_{21-22} p_{22-22}^{(n_4-1)} p_{22-F}.$$

Synchronism occurs only when $n_1 = n_3$ and expressions of cumulated probabilities become:

$$\prod_{i=1}^{n_1} p(x_i|s_{11}) \prod_{i=n_1+1}^{N} p(x_i|s_{22})$$

$$p_{11-11}^{(n_1-1)} p_{11-22} p_{22-22}^{(n_2-1)} p_{22-F}.$$

Using the product of eqs (2) and (3) as a recombination rule, we obtain a global probability that we must compare with the equivalent expression for the meta-HMM. In particular, taking (1) into account the contribution of transition probabilities to the global probability is:

$$q_1q_2q_3q_4(p_1p_4)^{(n_1-1)}(p_2p_4)^{(n_2-1)}(p_3/p_4)^{(n_3-1)}.$$

It is easy to check that products of emission pdf's of the upper and lower stream states (according to table I) are equal to the pdf's of the corresponding meta-state, namely:

$p(x\|s_{11}) = p(\xi\|s_1)p(\eta\|s_3)$
$p(x\|s_{12}) = p(\xi\|s_1)p(\eta\|s_4)$
$p(x\|s_{21}) = p(\xi\|s_2)p(\eta\|s_3)$
$p(x\|s_{22}) = p(\xi\|s_2)p(\eta\|s_4)$

Table II: Correspondance between emission pdf's of meta-states and multistream model states.

Table II shows that the mean vectors of the meta-states are build by appending mean vectors of the corresponding states according to table I. Also covariance matrices of the meta-states are block-diagonal ($d_u \times d_u$ and $d_l \times d_l$) with blocks corresponding to the states as in table I. This equivalence implies that upper and lower vectors are assumed independent since the emission probabilities of the meta-states factorize. In the training of the meta-HMM, mean vectors and covariances matrices will be constrained to meet this specific structure in order to preserve equivalence. Equivalence between transition probabilities is discussed in the next section.

3. TRANSITIONS

Equivalence between transition probabilities is less straightforward. Of course, it is of common practice to neglect transition probabilities in standard HMM models since their role has been shown of second order compared with emission probabilities. However, if equivalence between models is studied, the relations between transition probabilities is crucial since neglecting them could lead to a meaningless equivalent model.

We will show that there is no possible equivalence when enforcing equivalence together with the constitutive contraint of HMM which requires that the outgoing transitions of a state sum up to unity. This constraint will be neglected since the equivalent model is nothing else but a computational tool referred to as meta-HMM for that reason. However, constraints on the multistream model apply and must be replicated in this meta-HMM as shown below.

Comparing of (2) and (3) with the global probability of the meta-HMM yields successively in cases $n_1 > n_3$, $n_1 < n_3$ and $n_1 = n_3$ since the equivalence must hold for any n_i:

Case $n_1 > n_3$	$p_{12-12} = p_1p_4$ $p_{22-22} = p_2p_4$ $p_{11-11}/p_{12-12} = p_3/p_4$ $p_{11-12}p_{12-22}p_{22-F}/p_{12-12} = q_1q_2q_3q_4$
Case $n_1 < n_3$	$p_{11-11}p_{22-22}/p_{21-21} = p_1p_4$ $p_{22-22} = p_2p_4$ $p_{21-21}/p_{22-22} = p_3/p_4$ $p_{11-21}p_{21-22}p_{22-F}/p_{21-21} = q_1q_2q_3q_4$
Case $n_1 = n_3$	$p_{11-11} = p_1p_3$ $p_{22-22} = p_2p_4$ $p_{11-22}p_{22-F} = q_1q_2q_3q_4$

Table III: Set of equivalence constraints.

From these equations, it is easy to observe that the following equivalences can be derived for the self-loop transitions:

$p_{11-11} = p_1p_3$
$p_{22-22} = p_2p_4$
$p_{12-12} = p_1p_4$
$p_{21-21} = p_2p_3$

Table IV: Equivalence relations between self-loop transition probabilities.

For the inter-state transitions, the following relations must be satisfied:

$$p_{11-12}p_{12-22} = q_1q_3p_1p_4(q_2q_4/p_{22-F})$$

$$p_{11-21}p_{21-22} = q_1q_3p_2p_3(q_2q_4/p_{22-F})$$

$$p_{11-22} = q_1q_3(q_2q_4/p_{22-F})$$

A possible solution (not the only one) is

$p_{11-22} = q_1q_3$
$p_{11-12} = p_1q_3$
$p_{12-22} = q_1p_4$
$p_{11-21} = q_1p_3$
$p_{21-22} = p_2q_3$
$p_{22-F} = q_2q_4$

Table V: A possible set of equivalence relations between inter-state transition probabilities.

The solution is unique for the first four equations (Table IV) while several solutions are possible for the inter-state transition probabilities. For the one described in Table V, the physical interpretation is straightforward by considering the multistream model: p_{11-11} is the self-loop transition on s_{11} i.e the upper and lower parts of the feature vectors loop respectively on s_1 and s_3 while p_{11-12} is the combined transition of the upper part keeping looping on s_1 and the lower part moving from s_3 to s_4.

There is no solution which could simultaneously guarantee that all outgoing probabilities of a meta-state are summing up to 1 but this is not really required. On the other hand, it is mandatory that $p_i + q_i = 1$ for $i \in [1, ..., 4]$ in the multistream model.

4. TRAINING META-HMM'S

Any utterance has a corresponding meta-model built by concatenating all the meta-models of sub-units (here phonemes). Now any alignment of the corresponding feature vector sequences yields a global probability which is the product of all upper and lower stream emission probabilities and all transition probabilities to the a power equal to the number of uses of this transition. This means that the contribution of the transitions has the generic form

$$\cdots p^a_{11-11} p^b_{11-22} p^c_{11-21} p^d_{12-12} p^e_{11-12} p^f_{12-22} \cdots$$

Using the equivalence relations of Tables IV-V, this product can be rewritten in terms of multi-stream model transition probabilities p_i and q_i with the appropriate exponents: for instance, the exponents of p_1 and q_1 are respectively $m_1 = a + d + e$ and for q_1 it is $r_1 = b + c + f$. Maximizing the logarithm of the global probability under the constraints $p_i + q_i = 1$ (Lagrange multipliers) yields $p_1 = m_1/(m_1 + r_1)$ and $q_1 = r_1/(m_1 + r_1)$. Using the same approach for each state and converting the results back to the meta-HMM (using Tables IV-V again) yields update of the models and allows iteration of the Viterbi algorithm. Of course neglecting the meta-HMM transitions (i.e. assuming them all equal) corresponds to no multi-stream model.

5. CONCLUSIONS

The advantages of the multistream models have been demonstrated by others mainly for recognition of speech corrupted by band limited noise [1-3].

The goal of this paper is to describe an algorithm taking rigorously account of the equivalence between two models: multistream HMM and meta-HMM. Preliminary tests on a small database have demonstrated the importance of transition probabilities. Indeed, if training and recognition are conducted without transition probabilities, the recognition score at the phoneme level was for a baseline 1-stream model 72.5% against 73% for 1-stream with transition probabilities as expected. However, in 2-stream models we observed a degradation down to 40% without transition probabilities. The role of transitions is also enphasized if recognition uses probabilities updated during the training process but not used in the forced alignment. However, the size of the data base was not fully representative and systematic tests are conducted on TIMIT and will be reported at the conference.

6. REFERENCES

1. H. Bourlard, S. Dupont, H. Hermansky, N. Morgan, "Towards sub-band-based speech recognition," in *Proc. of European Signal Processing Conference*, (Trieste, Italy), pp. 1579–1582, Sept. 1996.

2. H. Bourlard, S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. of Intl. Conf. on Spoken Language Processing*, (Philadelphia, USA), pp. 422–425, Oct. 1996.

3. H. Bourlard, S. Dupont, C. Ris, "Multi-stream speech recognition," Tech. Rep. IDIAP-RR 96-07, IDIAP, Martigny, Switzerland, 1996.

4. S.Dupont, H.Bourlard, "Using Multiple Tiem Scales in a Multi-stream Speech Recognition System", *Proc. Eurospeech 1997*, vol 1, pp.3, Rhodes Greece, 1997