Speech Recognition with Factorial-HMM Syllabic Acoustic Models

Gianpaolo Coro[*°], *Francesco Cutugno*[*], *Fulvio Caropreso*[*]

[*] Department of Physics, University of Naples "Federico II", Naples, Italy
[°] ABLA 'beyond the voice' s.r.l., Milan, Italy
gianpaolo.coro@abla.it, cutugno@na.infn.it, ful77@libero.it

Abstract

Approaches in Automatic Speech Recognition based on classic acoustic models seem not to exploit all the information lying in a speech signal; furthermore decoding procedures have real time constraints preventing the system to achieve optimal alignment between acoustic models and signal. In this paper, we present an approach to speech recognition in which Factorial Hidden Markov Models (FHMM) are used as syllabic acoustic models. An alignment algorithm is used for unit decoding. As applicative domain we choose numbers (range 0-999,999) uttered in Italian. Syllabic accuracy in our model is 84.81%, correctness on numbers is 77.74%. Aim of the experiment is to show that the performances of FHMMs lie in the ability to retrieve the presence of two different temporal dynamics in a speech segments: the former with a quasi-segmental timing, the latter presenting a quasi-syllabic trend. Moreover, we evaluate a unit decoding process based on a dynamic programming algorithm in order to exploit the acoustic models performances at best.

Index Terms: Factorial HMM, speech recognition, syllabic acoustic models

1. Introduction

Many works from different scientific communities, ranging from neurobiology to experimental phonetics, suggest that speech signals contain information distributed on different time scales. Furthermore, it is necessary that more parallel cognitive functions operate a chunking on the unfolding of the information over time in order to process speech signals properly [1]. Humans seem to perform speech recognition successfully because of a partial parallelization process. The left-to-right speech stream is captured in a multilevel grid in which several linguistic analyses take place simultaneously.

Evidence of parallelized speech processing can be found in literature [1]. These ideas have rapidly influenced many research projects on Automatic Speech Recognizers (ASR) as well as led to the introduction of new concepts like syllabic pre-segmentation, word n-gram statistical combination, parallel and multi-scale speech coding [2] [3] [4] [14].

At the same time, Factorial HMMs (FHMM) [5] have been widely used in cases in which several, concurrent dynamics take place in an event production. Applications can be found especially in speaker separation tasks, where two or more voices overlap and a recognizer has to act on one of these [6].

In our opinion, speech production can be seen as a multi concurrent process, in which different time scale dynamics take place. A FHMM could then retrieve the presence of these behaviors on a syllable scale speech segment from which it separates a slow syllabic process from a fast phonetic one.

2. Motivation and related work

Greenberg's theory about the multi-granular nature of speech understanding [2] has been developed in two experiments where information coming from different time scales of analysis was integrated. Wu faces the problem of integrating two different recognizers [3], the former based on phonetic units, the latter based on the so called "half-syllable". Several experiments are made in the attempt to find the best way for integrating different scales of information. The best model is found using a mixture of a half syllables and phones as basic speech unit in word decoding. Wu's system, tested on clean speech for recognition of numbers ranging from 0 to 999, results in a 6% Word Error Rate. Furthermore, in several cases, Wu states that the two recognizers, constituting the baseline systems for performance comparison, produce complementary errors which can compensate each other in a multi-scale decision system.

A different approach for multi-granularity representation is made by Chang [4]. In this work, a "multi-tier model" is built, and speech is organized as a sequence of syllables, in contrast to the conventional "phonetic-segment-based organization" assumed in most ASR systems. It differs from conventional syllabic models as it represents single syllables as a set of acoustic "cues" instead of a succession of phonetic features. Acoustic cues refer to the phonetic structure of the syllable in terms of manner, place of articulation, vowels, etc., furthermore. some non-segmental information aiding the recognition of a syllable in a word are added. Word templates made up of a succession of previously defined features have been introduced and the possibility of mutation for these descriptions is associated to pronounce variability. Many other authors could be quoted following this examples, and all their works underline the importance of facing multi- granularity in speech processing as a crucial step for outperforming ASR design and implementation.

On the other side, recent work about Factorial HMMs has put in evidence their ability to separate multi-channel sources in cases where concurrent dynamics have well-defined statistical characteristics [6]. Furthermore FHMMs have been used in speech recognition, in which voice overlap had to be faced [7], or as acoustic models for phonemes classification [8]. In this scenario we choose to use such models in syllable recognition. The underlying idea is that in a syllable speech segment, two dynamics can be retrieved, one with a fast segmental trend, which goes through the "fine structure" of the signal portion, and the other with a slower rate related to larger syllabic temporal granularity. The two dynamics can be seen as two processes with a well independent nature, and a Factorial model can be used to let naturally emerge this dual phenomenon. A speech recognizer based on a set of FHMM syllabic acoustic models and its performances will be discussed in this paper. A comparison between standard HMMs and Factorial models performances supports our conclusions.

In addition, a novel approach to unit decoding is presented. It tries to find the best alignment of syllables to the signal after the utterance recording has stopped. It is not a real time algorithm, and it has a higher complexity compared to the most used algorithms but it results in high performances. We introduced this procedure in order to exploit our Factorial model at best, without loss of performance due to real time needs or beam search approximations. Performances of this decoding strategy will be presented in either the case of using a standard HMM acoustic model or a Factorial acoustic model.

3. Factorial HMMS

FHMMs, firstly introduced in [5], are HMMs whose state set can be decomposed in *L* subsets. Each subset evolves independently as a standard Markov chain and all subsets jointly contribute to the observable variables generation, as shown in Fig. 1.

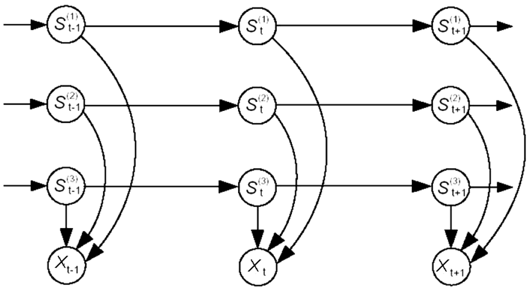

Figure 1: *Factorial HMM dynamic from* [5].

In a standard Hidden Markov Model, a sequence of observations $X = X_1, X_2,.., X_T$ is modelled by specifying a probabilistic relationship between the observations and a sequence of hidden states $S = S_1, S_2,.., S_T$ taken by a finite set of states of dimension *K*. Moreover the model assumes that observations are independent on each other and that each S_t is only dependent on S_{t-1}. HMM models are defined by the probability $P(S_{t-1} | S_t)$ of state succession, which is a $K \times K$ transition matrix and by the emission probabilities $P(X_t | S_t)$ which link the states to the observations. Such values can be calculated in many ways, in the case of continuous observation vectors, a gaussian mixture or a neural network can be used.

Factorial Hidden Markov Models expand the concept of HMM by representing a single state S_t as a collection of *M* states

$$S_t = S_t^{(1)}, S_t^{(2)},.., S_t^{(m)},.., S_t^{(M)} \qquad (1)$$

each of which can take on $K^{(m)}$ values (for simplicity we will assume $K^{(m)} = K$ for all m). So a FHMM consists of a state space which can be described by a $K^M \times K^M$ transition matrix. Such a system is equivalent to a HMM with K^M states, and all variables are allowed to interact arbitrarily. The processing complexity is obviously exponential in M. Interesting phenomena come out when constraints are introduced in the state transition matrix. As far as our application concerns, each state variable $S^{\{m\}}$ is allowed to evolve according to its own dynamic, so that:

$$P(S_t | S_{t-1}) = \prod_{m=1}^{M} P(S_t^{(m)} | S_{t-1}^{(m)}) \qquad (2)$$

Fig. 1 depicts this structure. The transition between states can be represented as *M* distinct $K \times K$ matrices.

As far as the emission probability of the observation X_t concerns, a gaussian distribution can be introduced, whose mean will depend on the $S_t^{(m)}$ states

$$\mu_t = \sum_{m=1}^{M} W^{(m)} S_t^{(m)} \qquad (3)$$

where each $W^{(m)}$ is the contribution of $S_t^{(m)}$ to the mean. The covariance matrix length depends on the X_t observation vector length.

$$P(X_t | S_t) \propto \exp(-\frac{1}{2}[(X_t - \mu_t)' C^{-1} (X_t - \mu_t)]) \qquad (4)$$

FHMMs have shown to be able to decompose automatically the state space into features differentiating multiple dynamics concurring in a single phenomenon. This is particularly efficient in cases in which the data are known to be generated from the interaction of multiple, loosely-coupled processes [5]. Our idea is that the multi-granular information in speech production, coming from syllabic and phonetic structure, may be thought as generated by overlapping processes with different time spanning, and a factorial model can catch these dynamics. The training can be able to associate different time-scale phenomena to different chains of the state set automatically. The layered nature of the model arises by only allowing transitions between states in the same layer. This division models processes with loosely coupled dynamics. Each layer operates similarly to a basic HMM but the probability of an observation at each time depends upon the current state in all the layers. In our work only two levels of chains are used, one for phonemes and another for syllables. This structure constitutes the acoustic model of the speech recognizer.

4. The decoding algorithm

In order to exploit our model at best, an efficient algorithm for syllable decoding had to be developed. Standard algorithms usually act in real time using dynamic programming methods and some approximations (as in the case of the beam search algorithm) with tha aim to reduce execution time. These procedure can introduce many errors as the recognition is strongly dependent from the left-to-right time processing. A more efficient procedure could try any possible alignment between words and signal, as it could retreat some decisions made during the left-to-right processing, and could try to shift the models backwards or forwards to achieve the best alignment and word separation. This procedure has been developed using dynamic programming. It does not act in real time because of our request to be independent from the signal runtime generation. This is undoubtedly an high complexity procedure, however we believe that this could drive acoustic models performance at best.

The main aim of the algorithm is to navigate the structure formed by the union of the language and acoustic model in order to maximize the probability *P(W|X)* for a sequence of units (syllables in our case) $W=w_1w_2...w_n$ given the observation sequence $X=X_1X_2...X_T$.

We could calculate *P(W|X)* as follows:

$$P(W \mid X) = P\left(w_1,\dots,w_m \middle| X_1^t\right) \cdot P\left(w_n \middle| w_m\right)^{\gamma} \cdot P\left(w_n \middle| X_{t+1}^T\right) \tag{5}$$

where w_n is the last unit if *W* is not empty and w_m is the preceding syllable in the sequence. $P(w_n|w_m)$ is the language model probability between w_m and w_n, γ is the language model weight, and *t* is the optimal time boundary between the units. Let's demonstrate that if *P(W|X)* is the optimal solution for the units alignment problem, then $P(w_1w_2...w_n\ |X_1^t)$ is the optimal solution for the problem of units alignment in the time interval [1,*t*], where *t* is the best first boundary for w_n. This is trivial in the fact that if there was another sequence $w'_1w'_2...w'_n$ for which $P(w'_1w'_2...w'_n|X_1^t)>P(w_1w_2...w_n\ |X_1^t)$ then it would be

$$P\left(w'_1...w'_k \middle| X_1^t\right) \cdot P\left(w_n \middle| w'_k\right)^{\gamma} \cdot P\left(w_n \middle| X_{t+1}^T\right) > P(w_1...w_m w_n | X_1^T) \tag{6}$$

against the hypothesis of *P(W|X)* to be the optimal solution for the problem. This discussion leads us to introduce the following recurrence relation for the solution *f(m,t)* to the subproblem of unit alignment in the time interval [1,*t*]

$$f(m,t) = \max\begin{cases} P\left(X_1^t \middle| m\right) \cdot \pi(m) \\ \max\limits_{\substack{1 \le t^* < t \\ n \in Syl}} \left\{ f(n,t^*) \cdot P(m|n)^{\gamma} \cdot P\left(X_{t^*+1}^t \middle| m\right) \right\} \end{cases} \tag{7}$$

where *Syl* is the set of all the units involved, *P(m|n)* is the probability of *n* and *m* unit concatenation, $P(X_1^t|m)$ is the likelihood of the model *m* to the observations $X_1X_2...X_T$, and $\pi(m)$ is the probability for *m* to be a starting unit for a sequence. Notice the dependency from $f(n,t^*)$, which is the best solution to the subproblem of units alignment till time instant t^*.

The optimal solution will be retrieved as follows

$$P(W \mid X) = \max_{m \in Syl} \{ f(m,T) \cdot E(m) \} \tag{8}$$

Where *E(m)* is the probability for the model *m* to be a plausible ending unit. Starting from this solution, a backtracking procedure produces the best alignment. The algorithm is also based on the calculation of the matrix *V*, which contains the likelihoods of a model *m* to all the intervals of observations.

$$V = \begin{pmatrix} P\left(X_1^1 \middle| m\right) & P\left(X_1^2 \middle| m\right) & \dots & P\left(X_1^{T-1} \middle| m\right) & P\left(X_1^T \middle| m\right) \\ -\infty & P\left(X_2^2 \middle| m\right) & \dots & P\left(X_2^{T-1} \middle| m\right) & P\left(X_2^T \middle| m\right) \\ \dots & \dots & \dots & \dots & \dots \\ -\infty & -\infty & \dots & -\infty & P\left(X_T^T \middle| m\right) \end{pmatrix} \tag{9}$$

The algorithm complexity is $O(T^2N^2C(V))$, were *C(V)* is the complexity of the likelihood calculations for a single model. If S is the number of states in the acoustic model, $C(V)=O(S^2T)$. This value can be reduced considering that using controlled (even if connected) speech, a single syllable can rarely have a maximum duration greater than a fixed values (500 ms in our case). At this point, the matrix *V* will get a band aspect which allows optimization about complexity issues.

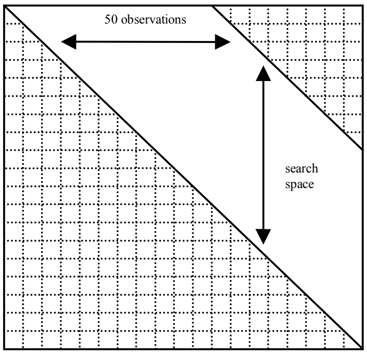

Figure 2: *Band matrix for complexity reduction based on assumptions about syllable length*

If observations are taken every 10 ms this means that we can calculate the likelihoods only on 50 observations intervals leaving to zero longer span probabilities. The complexity of this algorithm seems to be quite high for practical applications, especially if the utterance is too long, but it leads to an optimal alignment. Our tests have stated that, on an AMD 2800+ processor, the response is about 30 seconds, for a 3 seconds utterance, after the recording stops. Efforts should be fronted in the next future to improve this performance.

5. ASR

Our speech recognizer uses FHMM acoustic models and the above presented decoding algorithm for utterance recognition. Notice that we adopted a definition of syllable which is inspired by a phonological in which we can distinguish an onset, a nucleus and a *coda*. On the other hand, our approach is oriented to individuate in the energy temporal pattern particular regularities [9] with the consequence that our syllabification process can, in some case, deviate from the one usually defined by linguistic rules in the Italian phonological framework. Based on this assumption, syllabic units were obtained looking for energy islands separations. The language model has been weighted in order to give the best results. The same schema has been used to build up a reference ASR which employs standard HMM acoustic models. MFCC as defined in [10, (9):424-426] were used as features because of their fine grain structure. Acoustic models are not usually able to retrieve a phonetic inner dynamic from syllable length features, also because of the lower amount of information carried by the observations. The sequence of MFCC vectors stores information about the phonetic structure as well as coarticulation phenomena then it is foreseeable that, from these features, a factorial model could extract the syllabic dynamic presence too.

6. Results

In this section we will show results obtained from the section of SPEECON [11] corpus collecting numbers uttered by . The task involves numbers ranging form 0 to 999,999. For a motivation of this choice and for state-of-art achievements in this field see [12]. The corpus is formed by about 2000 utterances, with about 30 examples for each syllable training. As previously stated, MFCC are used as features in every system.

Table 1 presents a comparison in performances between syllable acoustic models built both with standard HMMs and Factorial HMMs.

	Accuracy	Correctness
Standard HMM	81.33%	89.11%
Factorial HMM	84.81%	94.30%

Table 1*: performances on syllable classification.*

Table 2 shows the mean of the *permanence in state* (number of auto-transitions) for each of the two layers of the FHMM acoustic model, calculated on all the models. Generally two processes with different speed are retrieved form the signal, this is evident especially when a syllable is long and rich of voiced phonemes.

Faster level	Slower Level
4.5016	7.8357

Table 2*: mean permanence in state for FHMM layers calculated on all the models*.

Table 3 shows a comparison between three systems: the first has been built with HTK [13]. It uses syllabic acoustic models, and a Token Passing algorithm for word decoding. The second system makes use of syllabic acoustic models with a standard HMM structure and of the new algorithm (cfr. section 4) for decoding. The last system integrates FHMM acoustic models and our decoding algorithm for the decoding procedure.

	Correctness
HTK	62.14 %
ASR with Standard HMM	74.09 %
ASR withFactorial HMM	77.74%

Table 3: *performance comparison between ASRs on the recognition of a single number in the range 0-999,999*.

7. Discussion

Performances in Table 1 show that Factorial HMMs can achieve an higher performance on syllable classification.

Table 2 shows that concurrent dynamics, different in speed, are actually detected. It is particularly evident when syllables have a structure which is far from a single voiced speech sound. Table 3 shows that, even if the complexity of the novel decoding algorithm is high, it produces best results than a classic recogniser based on a Token Passing strategy. FHMMs performance are even more emphasized by this exact decoding procedure.

8. Conclusions

We have investigated the application of Factorial HMMs to acoustic modeling in automatic speech recognition. An high complexity decoding procedure has been used for word decoding to exploit at best models performances on a task of numbers recognition. Results have supported the idea that a syllable stores slower and faster dynamics. Factorial HMM models can retrieve the presence of these two trends. The decoding strategy still has an high complexity, even if responses are given in less than 30 seconds for an utterance of 3 seconds on a standard PC, but it seems to exploit FHMMs power. Future work will regard the investigation of new ways to reduce Factorial models recognition complexity without heavily affecting performances. New ways for detecting inter-syllabic or inter-word dynamics will also be investigated, with the aim to exploit information coming from different time scales, adding higher levels of knowledge sources.

9. References

[1] Poeppel, D., “The Analysis of Speech in Different Temporal Integration Windows: Cerebral Lateralization as ’Asymmetric Sampling in Time’”, Speech Communication, 41:245-255, 2003.

[2] Greenberg, S., “Understanding Speech Understanding: Towards a Unified Theory of Speech Perception”, ESCA Workshop on Auditory Basis of Speech Perception, 1–8, 1996.

[3] Wu., S.L., “Incorporating Information from Syllable-Length Time Scales into Automatic Speech Recognition”, Ph.D. Thesis, University of California, Berkeley, 1998.

[4] Chang, S., “A Syllable, Articulatory-Feature and Stress-Accent model of Speech Recognition”, Ph.D. Thesis, University of California, Berkeley, 2002.

[5] M. Jordan, Z. Ghahramani, “Factorial Hidden Markov Models”, Machine Learning, 29:245–273, 1997.

[6] Reyes-Gomez, M., Raj, B., and Ellis, D., “Multi-channel source separation by factorial HMMs”, In Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Hong Kong, 2003.

[7] Deoras, A.N., Hasegawa-Johnson, M., “A Factorial HMM Approach to Simultaneous Recognition of Isolated Digits Spoken by Multiple Talkers on one Audio Channel”, in Proc. of Interspeech’04, 2093-2096, 2004.

[8] Logan, B.T., Moreno, P.J., “Factorial HMMs for acoustic modeling”, In Proc. ICASSP’98, 813-816, 1998.

[9] Cutugno F., Passaro G., Petrillo M. “Sillabificazione fonologica e sillabificazione fonetica”, in Albano Leoni F., Sornicola R., Stenta Krosbakken E., Stromboli C. (eds), Atti del XXXIII, Congresso della Società di Linguistica Italiana, Bulzoni Roma, 205-232, 2001.

[10]Huang, X., Acero, A., Hon, H., “Spoken Language Processing”, Prentice Hall, 2001.

[11] SPEECON corpus “http://www.elda.org/article10.html”

[12] Rahim M., Riccardi G., Saul L. Wright J., Buntschuh B., Gorin A. , “Robust numeric recognition in spoken language dialogue”, Speech Communication, 34:195-212, 2001.

[13]Hidden Markov Model Toolbox Kit (HTK) http://htk.eng.cam.ac.uk/

[14]Deng, L.,Sun, D., “A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features”, Journal of the Acoustical Society of America, 95:2702-2719, 1994.