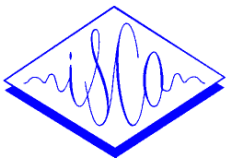

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

Sound duration modelling and time-variable speaking rate in a speech recognition system

Nelly Suaudeau

IRISA/INRIA
Campus de Beaulieu
35042 Rennes France

Régine André-Obrecht

IRIT/CNRS URA 1399
Université Paul Sabatier
31062 Toulouse France

ABSTRACT

Among the features extracted from the speech signal, it's clear that someone are directly dependent on the elementary acoustic level whereas the others depend on the suprasegmental level such as the phonetic level. A major deficiency of a standard HMM is that it takes into account uniformly the informations. In this paper, we try to resolve this problem using the Two Level HMM which introduces the features with respect to their informative contents, either on the elementary acoustic level or on the phonetical level. Namely, the incorporation of global sound durations is explored. More, since variations in speaking rate affect sound durations, we propose to appropriatly adapt the sound duration pdf parameters. Experiments on french number database show that such an explicit introduction of prosodic parameters improves the recognition accuracy.

Keywords: *Speech recognition; Markov modelling; Prosodic informations; Sound duration; Speaking rate.*

1 INTRODUCTION

In speech signal, the acoustic expression of an underlying phonetic unit such as a phoneme corresponds to a succession of quasi-stationary events, each event exhibiting its specific spectral characteristics. The Hidden Markov Models (HMM) are succesfully performed in speech recognition systems, exploiting the instantaneous spectral features of the acoustic signal. Each word is first decomposed as a sequence of subword linguistic units (allophones, phonemes, ...). A markovian source is then associated to each of these elementary linguistic units, the result being a global acoustic network. This procedure leads to an accurate representation of speech acoustic structure. Nevertheless, it is not sufficient to model prosodic parameters (energy, duration) that are important cues too. Now, in order to improve the recognition performance, not only the spectral aspects but also the prosodic informations need to be processed by the recognition system.

Some solutions are proposed to take sound duration into account: boundaries of sound duration are imposed ([2]), duration laws may be associated to every state with the state corresponding to a phone ([3]) or to a subphone ([4]). However, a trade-off between an efficient acoustic modelling and a realistic prosodic one makes these appoaches not satisfactory. Indeed, all the observations are introduced at a same hierarchical level with only one index scale. Now, unlike the spectral features which are directly dependent on the elementary acoustic level (or infraphonemic level), prosodic parameters are related to the phonetic suprasegmental level. Moreover, the speaking rate remains unknown.

Therefore, for representing more adequatly the totality of the pertinent informations extracted from the signal, we have studied an alternative model, the two-level Hidden Markov Model (TLHMM) ([5]): the traditionnal **hierarchical** structure of a classical HMM is retained. This model specificity lies upon new hypothesis formulations concerning the output probability distibutions of the prosodic components of the observations. In order to validate such an approach, we use the TLHMM to model the global sound duration. Furthermore, since speaking rate is one of the factor recognised to affect sound duration, the parameters of the duration laws are adapted either during the recognition phase or in a post-processing step.

2 THE TLHMM

The Two-Level HMM results from an adaptation of a standard HMM, built according to a traditional hierarchical structure : an intermediate level provides a phonetical description of the authorized sentences and each elementary unit (subword unit) is associated to an elementary acoustic model. From these informations, the compiler gives the global acoustic network. The specificity lies upon new hypothesis formulation concerning the output probability distributions of the observations. The observation vectors remain assumed to be generated by a state sequence at the acoustic level, but each one is divided into two components according to its informative content : phonetic cues are distinguished from acoustic ones.

More specifically, this probabilistic model is defined from three stochastic processes (X_t, Y_t, P_t) :

- $(X_t)_{t\geq 1}$ is a first order markovian process. Due to the hierarchical network topology, it is possible to determine, given a state sequence $(X_t)_{t=\overline{1,T}}$ (extracted from the acoustic level), its corresponding sequence of subword units, characterized by the process $(W_\tau)_{\tau=\overline{1,\epsilon}} = (\phi_{j_\tau}, \kappa_\tau)$. ϕ_{j_τ} denotes the symbol of the τth subword unit. The integer κ_τ is the time

index t of the first state belonging to the τth subword unit.

- $(Y_t)_{t\geq 1}$ is an observable process representing the acoustic observations. Conditionally to the state sequence, the acoustic observations are assumed to be independent of each others as with the HMM.
- $(P_t)_{t\geq 1}$ is an observable process whose outputs are the phonetic observations. We assume that:
  - conditionally to the state sequence, this process is independent of the acoustic observations $(Y_t)_{t\geq 1}$,
  - each phonetic observation P_t is not directly tied to the process $(X_t)_{t\geq 1}$ but it is a probabilistic function of the phonetic process $(W_\tau)_{\tau=\overline{1,\epsilon}}$,
  - two consecutive phonetic observations are correlated with each other if and only if they are emitted within a same subword unit and they jointly depend on the current subword unit.

By summing up or averaging, ..., the phonetic observations $P_{\kappa_\tau}, \ldots, P_{\kappa_{\tau+1}-1}$, a global value G_τ is determined. This global feature G_τ is then explicitly modelled with a probability density function (pdf) ϑ_{j_τ} associated to the current subword unit ϕ_{j_τ}. Eventually, complementary probabilistic assumptions are added in order to take into account the way this global value is shared among the segments aligned to the phonetic unit: a probabilistic function $coef(..)$ is used.

$$Pr[\underbrace{p_{\kappa_\tau}, \ldots, p_{\kappa_{\tau+1}-1}}_{\rightarrow g_\tau} \mid \phi_{j_\tau}) = \vartheta_{j_\tau}(g_\tau) \times coef(g_\tau, \kappa_\tau, \kappa_{\tau+1})$$

Thus, the TLHMM permits to introduce the observations at two distinct hierarchical levels in terms of their informative content and every level process the observations at its own cadence.

In order to use the TLHMM for speech recognition, we have to solve the recognition problem and to reestimate the model parameters. Due to the particular dependency proprieties of the phonetic observations, we can not use classical facilities (Baum algorithm, Viterbi algorithm). Therefore, the recognition and parameter estimation algorithms must be extended. A comprehensive treatment of the recognition and parameter estimation procedures derived for the TLHMM may be found in ([5]).

3 BASELINE SYSTEM

Our baseline recognition system includes two modules: the acoustic pre-processing and the linguistic decoder. It differs from a standard frame-based HMM in that:

- measurements are made on variable-duration segments rather than fixed-length frames,
- a TLHMM is introduced in the linguistic decoder so as to explicitly model sound global duration.

During the acoustic pre-processing, an automatic segmentation algorithm ([4]) isolates the homogeneous parts of signal. Then, a spectral analysis is performed upon each variable length segment to provide one observation vector which is composed of conventional spectral parameters, and of a supplementary element, the segment length.

In a first experiment, the linguistic decoder is based on a hierarchical HMM : each word is described as a string of pseudo-diphones, one for each stationary part of phonemes and one for each transition between phonemes. Then, the topologies of the elementary Markov sources associated with these phonetic units are appropriatly choosen with respect to the acoustic segmental structure ([4]) ; the variable segment length is considered as a component of the observation vector. The association of the segmental preprocessing with such a standard HMM permits a first improvement: compared to the classical recognition systems the temporal structure of speech is better represented.

Yet, since the duration information remains integrated at the state level, such an approach is still insufficient the obtention of reasonable duration at the phonetic level isn't guaranted. Nevertheless, one can notice that by summing up the lengths of adjacent segments generated in a given subword unit, the calculated quantity (denoted d_τ) represents the subword unit duration ; more, despite the fact that numerous factors such as syllabic stress, adjacent sounds, ..., contribute to the phonemic duration variability statistical studies have demonstrated that a sound duration is hightly correlated with its identity.

Therefore, we have introduced this subword duration by implementing the TLHMM instead of the previous HMM to obtain our baseline system : a subword unit is still a pseudo-diphone, the segment lengths are considered as phonetic observations whereas the cepstral parameters give the set of acoustic observations. Furthermore, in order to explicitly model the subword unit duration d_τ, probability laws are associated with each subword. The distribution of the pseudo-diphone duration into segment lengths is considered as uniform, so the function $coef(.., .., ..)$ represents a multinomial distribution.

4 SPEAKING RATE MODELLING

When a speaker alters articulation times, one of the major effects is the reorganization of the speech temporal structure. So, the speaking rate is a source of important variations in signal and the magnitude of this phenomenon varies according to the phonemes. The fact that vowels are recognised to be more elastic than consonants is one such example. However, for simplicity, most of the studies assume that speaking rate *uniformly* affects the phonemic durations.

In our baseline system, the speaking rate isn't taken into account. In order to improve it, we are interested in two different methods to incorporate this information. The training procedure of the TLHMM parameters isn't modified, the speaking rate remains unconsidered during the learning phase. We propose :

- an adaptation of the sound duration pdf parameters with respect to the speaking rate during the recognition phase,
- a post processing including this speaking rate information to correct a first recognition hypothesis.

4.1 During the recognition process

The speaking rate model

The basic idea is that the speaking rate of each utterance can be represented by means of a factor denoted *eloc* and that pseudo-diphone durations depend on this factor *eloc* and standard statistical characteristics.

More precisely, let an utterance of which the phonetic realization is $\phi_1, \ldots, \phi_\tau, \ldots$, the speaking rate model is written as:

$$\begin{cases} eloc(\tau) &= eloc(\tau-1) + w(\tau) \\ d_{mes}(\tau) &= d_{mean}(\phi_\tau) \times eloc(\tau) + v(\tau) \\ eloc(0) &= 1 + w(0) \end{cases}$$

where:

- $eloc(\tau)$ represents the speaking rate factor and it may vary in course of time,
- $d_{mes}(\tau)$ is the measured duration of the τth phonetic unit,
- $d_{mean}(\phi_\tau)$ is the mean duration of the phoneme ϕ_τ (standard estimate),
- $w(\tau)$ and $v(\tau)$ are white Gaussian noise processes whose variances are respectively $\sigma_w^2(\tau)$ and $\sigma_v^2(\tau)$.

As the speaking rate decreases (resp: increases), the factor $eloc(\tau)$ becomes greater (resp: smaller) than 1 and it appropriety weights the standard duration mean of the τth phoneme, providing by this way a more likely duration parameter.

In our experiment, short sentences are performed, so we assume that the speaking rate depends essentially on both the speaker and the sentence. To take this specificity into account, the initial value $eloc(0)$ is a random variable with mean value 1 and variance $\sigma_w^2(0)$. We also assume that the speaking rate influence the observed durations : those effects are supposed in proportion to the means of the sound durations and the variance $\sigma_v^2(\tau)$ of the measurement noise is postulated to correspond to the phoneme ϕ_τ standard variance.

The estimation of the speaking rate model

A Kalman filter permits to recursively compute an estimate $\widehat{eloc}(\tau+1)$ given the past measured durations. At each time τ:

- using the predicted parameter, $\widehat{eloc}(\tau)$, a predicted value $\widehat{d_{mes}}(\tau)$ of the next measurement is computed,
- an innovation, the difference between this quantity and the measurement $d_{mes}(\tau)$, is evaluated,
- the innovation, the previous estimate of the speaking rate factor, the filter gain lead to the new estimate $\widehat{eloc}(\tau+1)$.

Adaptation of the TLHMM recognition algorithm

In the recursive equations of the extended Viterbi algorithm, the probability laws of the sound duration appear only when $(X_{t'})_{t' \leq t}$, a partial state sequence, is considered under the condition $t = \kappa_{\tau+1}$. At this time, the Kalman filter provides an estimation of the speaking rate factor from the past duration observations and the phonetic alignement corresponding to this path. The mean parameter of the pdf associated with ϕ_τ are adjusted according to the speaking rate factor: $d_{mean}(\phi_\tau)$ is substituded by $d_{mean}(\phi_\tau) \times \widehat{eloc}(\tau)$. The Viterbi equations may be implemented.

4.2 In a post-processing step

The previous approach has some limitations. During the estimation of the correcting sound duration factor, only a partial observation sequence is exploited ; furthermore the speaking rate of an utterance is reflected in syllable duration. Therefore, rather than deducing the speaking rate information from solely analysing sound durations, it seems better to obtain it from a larger context such as the syllabic unit.

The speaking rate model

The speaking rate is represented through a quantity $Syll_{mean}$, the syllabic mean duration. This quantity value is measured after identifying the whole sentence. To represent the correlation between duration and speaking rate we propose the following model:

$$\begin{cases} Syll_{mean} &= (\sum_{\tau=1}^{\epsilon} d_{mes}(\tau))/S \\ d_{mes}(\tau) &= \alpha(\phi_\tau) \times Syll_{mean} + \beta(\phi_\tau) + v(\tau) \end{cases}$$

where:

- $\phi_1, \ldots, \phi_\tau, \ldots, \phi_\epsilon$ is the recognized subword unit sequence,
- $d_{mes}(\tau)$ is the observed duration of the τth unit,
- S is the number of syllable,
- $(\alpha(\phi_\tau), \beta(\phi_\tau))$ is the linear regression coefficient vector associated with ϕ_τ,
- $v(\tau)$ is an error term.

Estimation of the model

Using such a model implies the estimation of the additional $(\alpha(\phi), \beta(\phi))$ parameters. Their values are calculated during the last iteration of the learning procedure. As the TLHMM training is based on a Viterbi procedure, for each training utterance, the most likely phonetic alignement (ie the Viterbi best path) is available. Then, we can recover the syllabic mean duration and, for each realization of the phoneme ϕ, its correlated observed duration. These measured features serve to define a data vector

$\omega^i(\phi) = (d^i_{mes}(\phi), Syll^i_{mean})$. Given the $(\omega^i(\phi))_i$ training data set, the estimation problem is performed using a MSE criterion, by minimizing with respect to parameters $(\alpha(\phi), \beta(\phi))$ the function:

$$\sum_i (d^i_{mes}(\phi) - \alpha(\phi) \times Syll^i_{mean} - \beta(\phi))^2$$

Recognition method

The speaking rate is deduced from the whole utterance, so an entire sentence hypothesis is required. Hence, this model can operate only in a post-recognition phase. We propose to apply a procedure based on a multi-pass search algorithm which consists in several successive steps:

- the standard TLHMM provides a preliminary sentence hypothesis, the most likely phonetic transcription. Either this solution is true, thereby the derived syllabic mean duration is exact ; or the recognition is unaccurate but the utterance has then a high probability of being partially recognized (in most error cases, only few phonemes are substituted, inserted or ommitted). As a result, we take advantage of the fact that the exact number of syllables in the token sentence remains near the measured ones to deduce 2N+1 very likely values for $Syll_{mean}$, denoted S^k :

$$S^k = (\sum_{\tau=1}^{\epsilon} d_{mes}(\tau))/(S+k)$$

  where $k \in \{0, \pm 1, \ldots, \pm N\}$. S represents the number of syllables included in the recognized sentence.

- for each speaking rate hypothesis, the statistic duration parameters in the TLHMM are adapted according to our current speaking rate model :

$$d_{mean}(\tau) = \alpha(\phi_\tau) \times S^k + \beta(\phi_\tau)$$

- using the extanded Viterbi procedure, every adapted TLHMM performs a new recognition, (2N+1) new propositions of the utterance, one per each S^k value, are proposed,

- to select one of these propositions, it is advisable to reject a sentence hypothesis if its a posteriori mean syllabic duration dismatches with those a priori postulated. A revised score is then calculated

5 EXPERIMENTS

The different techniques discussed above are evaluated on the French numbers (0 to 999) recognition task, using the number speech data-base from the CNET. This data-base, which contains about 5000 utterances recorded from 72 speakers, is split into two parts: one part for training and the rest for testing. Our recognition experiments are speaker independent. Moreover, the acoustic parameters are composed of the first 8 Mel frequency cepstral coefficients (MFCC), the first 4 time derivatives of the MFCC, the signal energy, the time derivative of energy and the segment length l. The developed modelizations use pseudo-diphones as subword units and the same global network (generated by the CNET network compiler) serves to implement the HMM and the TLHMM shemes as described in section 3.

Since a previous study ([5]) has demonstrated that the Inverse Gaussian distributions afford simple reestimation procedures and approximate accuratly sound durations these probabilistic laws are retained to model pseudo-diphone durations in the TLHMM framework. The more classical Gaussian functions are also tested. Results indicate that this new duration modelling improves the recognition rate from 94.6 % with standard HMM to 95.2 % with Gaussian laws and 95.35 % with Inverse Gaussian ones. We can notice that the Inverse Gaussian distributions, having positive supports, obtain the best recognition rate, which confirms that the performance improvement depends on the adequation between the speech reality and the choosen pdf.

Next, the speaking rate adaptation is experimented for the case in which the TLHMM models pseudo-diphone durations using Inverse Gaussian distributions. Our two proposed alternatives (section 4) permit to correct some mistakes as syllable insertions or omissions. Rather, both methods lead to a same recognition rate of 95.65 %. Yet only relatively small improvements are achieved: it may be explained by the fact that the considered sentences are short thereby exhibiting little speaking rate variations Consequently, as long as our recognition task is limited these preliminary results are interesting. They validate our approaches and suggest that a speaking rate adaptation presents potential benefits.

These experimental results show that an efficient combination between an accurate acoustic modelling and a realistic prosodic ones is essential in a speech recognition strategy to obtain a better recognition. The focus of future extensions to this work can take two basic directions First, it is desirable to test the proposed methods, namely this using a Kalman filter, on continuous speech databases Second, the integration of others supra-segmental informations such as energy and pitch is also of interest and for this purpose our methods seem to be especially suitable.

References

[1] F. Brugnara, R. De Mori, D. Giuliani, M. Omologo, A family of parallel Hidden Markov Models. *ICASSP 92*, 1992.

[2] Hung-yan Gu, Chiu-yu Tseng, Lin-shan Lee, Isolated-Utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations. *IEEE ASSP*, vol.39, no8, 1991.

[3] S.E. Levinson, Continuously variable duration hidden Markov Models for automatic speech recognition. *Computer Speech and Language*, Vol 1, No 1, pp 29-46, 1986.

[4] R. André-Obrecht, Reconnaissance automatique de parole à partir de segments acoustiques et de modèles de Markov cachés *18ièmes JEP*, Montréal, June 1990.

[5] N. Suaudeau, R. André-Obrecht, B. Delyon, Modélisation de la durée globale d'un son dans un modèle de Markov caché application à la reconnaissance des nombres. *19ièmes JEP*, 1992